Contents

Introduction

Computational neuroscience is an approach to understanding the information content of neural signals by modeling the nervous system at many different structural scales, including the biophysical, the circuit, and the systems levels. Theoretical analysis and computational modeling are important tools for characterizing what nervous systems do, determining how they function, and understanding why they operate in particular ways. The questions what, how, and why are addressed by descriptive, mechanistic, and interpretive models, each of which we discuss in the following chapters.

Definition 0.1. *Descriptive models* summarize large amounts of experimental data compactly yet accurately, thereby characterizing what neurons and neural circuits do.

Definition 0.2. *Mechanistic models* address the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry.

Definition 0.3. *Interpretive models* use computational and information-theoretic principles to explore the behavioral and cognitive significance of various aspects of nervous system function, addressing the question of why nervous systems operate as they do.

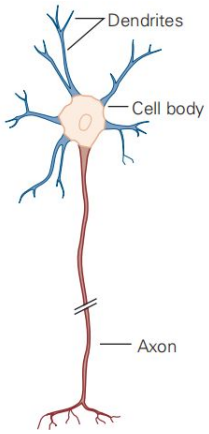
Chapter 1

Neural Encoding I: Firing Rates and Spike Statistics

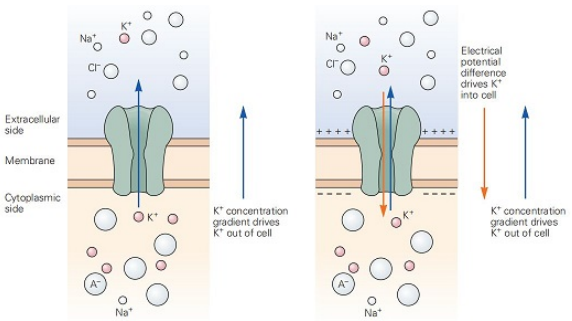
1.1 Introduction

1.1.1 Properties of Neurons

Remark 1.1. *Neurons* are highly specialized for generating electrical signals in response to chemical and other inputs, and transmitting them to other cells. *Dendrites* receive information inputs from other neurons. *Axons* carry the neuronal output to other cells. The cell body of a neuron is also called the *soma*.



Remark 1.2. *Ion channels* control the flow of ions across the cell membrane by opening and closing in response to voltage changes and to both internal and external signals.



Definition 1.1. The difference in electrical potential between the interior of a neuron and the surrounding extracellular medium is called the *membrane potential*.

Definition 1.2. Every cell, including a neuron, maintains a certain difference in the electrical potential on either side of the plasma membrane when the cell is at rest. This is called the *resting membrane potential*.

Remark 1.3. Under resting conditions, the potential inside the cell membrane is negative, outside the cell membrane is positive, and the cell is said to be *polarized*. *Ion pumps* located in the cell membrane maintain concentration gradients that support this membrane potential difference.

Definition 1.3. An *action potential* is the characteristic electrical pulses or, more simply, spikes that can travel down nerve fibers.

Definition 1.4. Current in the form of positively charged ions flowing out of the cell (or negatively charged ions flowing into the cell) through open channels makes the membrane potential more negative, a process called *hyperpolarization*.

Definition 1.5. Current flowing into the cell changes the membrane potential to less negative or even positive values. This is called *depolarization*.

Remark 1.4. If a neuron is depolarized sufficiently to raise the membrane potential above a threshold level, a positive feedback process is initiated, and the neuron generates an action potential.

Definition 1.6 (Refractory Period)**.** For a few milliseconds just after an action potential has been fired, it may be virtually impossible to initiate another spike. This is called the *absolute refractory period*. For a longer interval known as the *relative refractory period*, lasting up to tens of milliseconds after a spike, it is more difficult to evoke an action potential.

Remark 1.5. The absolute refractory period and relative refractory period are two basic phenomena in the process of neural response.

Remark 1.6. Action potentials are the only form of membrane potential fluctuation that can propagate over large

distances. They are regenerated actively along axon processes and can travel rapidly over large distances without attenuation.

1.1.2 Recording Neuronal Responses

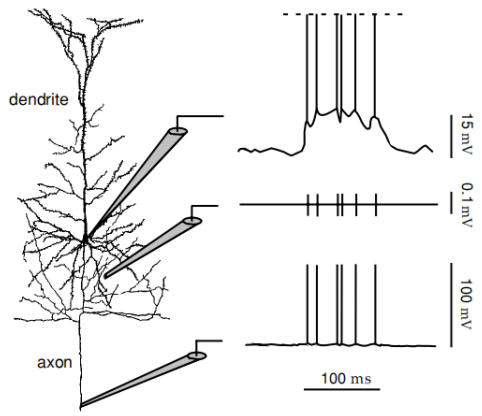
Rule 1.7. Membrane potentials are measured intracellularly by connecting a hollow glass electrode filled with a conducting electrolyte to a neuron, and comparing the potential it records with that of a reference electrode placed in the extracellular medium.

(i) *Intracellular recordings* are made either with sharp electrodes inserted through the membrane into the cell, or patch electrodes that have broader tips and are sealed tightly to the surface of the membrane. After the patch electrode seals, the membrane beneath its tip is either broken or perforated, providing electrical contact with the interior of the cell.

(ii) *Extracellular recordings* placed an electrode near a neuron but it does not penetrate the cell membrane.

Remark 1.7. Extracellular recordings can reveal the action potentials fired by a neuron, but not its subthreshold membrane potentials, and are typically used for in vivo experiments. Intracellular recording is more commonly used for in vitro preparations.

Example 1.8. Three simulated recordings from a neuron.



(i) The top trace represents a recording from an intracellular electrode connected to the soma of the neuron. The height of the action potentials has been clipped to show the subthreshold membrane potential more clearly. (ii) The middle trace is a simulated extracellular recording. (iii) The bottom trace represents a recording from an intracellular electrode connected to the axon some distance away from the soma.

1.1.3 From Stimulus to Response

Remark 1.8. Neurons typically respond by producing complex spike sequences that reflect both the intrinsic dynamics of the neuron and the temporal characteristics of the stimulus.

Definition 1.9. *Neural encoding* refers to the map from stimulus to response.

Example 1.10. We can catalog how neurons respond to a wide variety of stimuli, and then construct models that attempt to predict responses to other stimuli.

Definition 1.11. *Neural decoding* refers to the reverse map, from response to stimulus.

Remark 1.9. The complexity and trial-to-trial variability of action potential sequences make it unlikely that we can describe and predict the timing of each spike deterministically. Instead, we seek a model that can account for the probabilities that different spike sequences are evoked by a specific stimulus.

1.2 Spike Trains and Firing Rate

1.2.1 Firing Rate

Assumption 1.12. Action potentials are typically treated as identical stereotyped events in neural encoding studies ignoring in duration, amplitude, and shape. We ignore the brief duration of an action potential (about 1 ms), an action potential sequence can be characterized simply by a list of the times when spikes occurred.

Notation 1. Typically, a *spike train* that start at time 0 and end at time T can be characterized simply by $\{t_i\}_{i=1}^n$, a list of the times when spikes occurred, where $0 \le t_i \le T$ for all i and n represents the number of spikes in this spike train.

Definition 1.13. A *neural response function* $\rho(t)$ shows whether a spike is fired at time t, which can be represent as a sum of infinitesimally narrow, idealized spikes in the form of Dirac δ functions

$$\rho(t) = \sum_{i=1}^{n} \delta(t - t_i), \tag{1.1}$$

where $\{t_i\}_{i=1}^n$ is a spike train.

Lemma 1.14. For any well-behaved function $h(t)$, δ function satisfies

$$\int \delta(t - \tau)h(\tau)d\tau = h(t), \tag{1.2}$$

which provided that the limits of the integral surround the point t (if they do not, the integral is 0).

Proof. By the definition of δ function, we have

$$\lim_{\epsilon \to 0} \int_{t-\epsilon}^{t+\epsilon} \delta(t - \tau)h(\tau)d\tau = h(t) \int_{t-\epsilon}^{t+\epsilon} \delta(t - \tau)d\tau = h(t).$$

where the second step follows from the integral mean value theorem. □

Remark 1.10. $\rho(t)$ is used to re-express sums over spikes as integrals over time.

Theorem 1.15. For any well-behaved function $h(t)$, we can write

$$\sum_{i=1}^{n} h(t - t_i) = \int_{-\infty}^{\infty} h(\tau)\rho(t - \tau)d\tau, \tag{1.3}$$

where the integral is over the duration of the trial.

Proof. The equation directly follows from Equation 1.2. □

Proposition 1.16. The number of spikes n on a trial satisfies

$$n = \int_0^T \rho(\tau) d\tau.$$

Definition 1.17. The *spike-count rate* r of a spike train is obtained by counting the number of action potentials that appear during a trial and dividing by the duration of the trial,

$$r = \frac{n}{T} = \frac{1}{T}\int_0^T \rho(\tau) d\tau, \tag{1.4}$$

which indicates that the spike-count rate is the time average of the neural response function over the duration of the trial.

Remark 1.11. The spike-count rate can be determined from a single trial, but at the expense of losing all temporal resolution about variations in the neural response during the course of the trial. A time-dependent firing rate can be defined by counting spikes over short time intervals, but this can no longer be computed from a single trial, see Example 1.18.

Example 1.18. We can define the firing rate at time t during a trial by counting all the spikes that occurred between times t and $t + \Delta t$, for some small interval Δt, and dividing this count by Δt. However, for small Δt, which allows for high temporal resolution, the result of the spike count on any given trial is apt to be either 0 or 1, giving only two possible firing-rate values. The solution to this problem is to average over multiple trials.

Notation 2. We use angle brackets, $\langle\ \rangle$, to denote averages over trials that use the same stimulus.

Definition 1.19. The *time-dependent firing rate* $\mathrm{r}(t)$ is the the average number of spikes (averaged over trials) appearing during a short interval between times t and $t + \Delta t$, divided by the duration of the interval,

$$\mathrm{r(t)} = \frac{1}{\Delta \mathrm{t}}\int_{\mathrm{t}}^{\mathrm{t}+\Delta \mathrm{t}} \langle \rho(\tau) \rangle \mathrm{d}\tau, \tag{1.5}$$

where $\langle \rho(t) \rangle$ is the trial-averaged neural response function.

Notation 3. We use the notation $\mathrm{r}(t)$ opposed to r for the spike-count rate, and use the term "firing rate" without any modifiers, we mean $\mathrm{r}(t)$.

Remark 1.12. Formally, the limit $\Delta t \to 0$ should be taken on the right side of Equation 1.5, but, in extracting a time-dependent firing rate from data, the value of Δt must be large enough so there are sufficient numbers of spikes within the interval defining $\mathrm{r}(t)$ to obtain a reliable estimate of the average.

Proposition 1.20. For sufficiently small Δt, $r(t)\Delta t$ is the probability of a spike occurring during a short interval of duration Δt around the time t.

Proof. For sufficiently small Δt, $\mathrm{r(t)\Delta t}$ is the average number of spikes occurring between times t and $t + \Delta t$ over multiple trials. The average number of spikes over a longer time interval is given by the integral of $\mathrm{r(t)}$ over that interval. If Δt is small, there will never be more than one spike within the interval between t and $t + \Delta t$ on any given trial. This means that $\mathrm{r(t)\Delta t}$ is also the fraction of trials on which a spike occurred between those times. Equivalently, $\mathrm{r(t)\Delta t}$ is the probability that a spike occurs during this time interval □

Theorem 1.21. For any function h, we can replace the trial-averaged neural response function with the firing rate $\mathrm{r}(t)$ within any well-behaved integral.

$$\int h(\tau)\langle \rho(t-\tau)\rangle d\tau = \int h(\tau)\mathrm{r(t}-\tau)\mathrm{d}\tau. \tag{1.6}$$

Remark 1.13. Equation 1.6 establishes an important relationship between the average neural response function and the firing rate, the two are equivalent when used inside integrals. And provides another interpretation of $\mathrm{r}(t)$ as the trial-averaged density of spikes along the time axis.

Definition 1.22. The *average firing rate* $\langle r \rangle$ is the spike-count firing rate to be averaged over trials, that is,

$$\langle r \rangle = \left\langle \frac{n}{T} \right\rangle = \frac{\langle n \rangle}{T}, \tag{1.7}$$

where $\langle n \rangle$ is the trial-averaged number of the spike in the trial, T is the time period of the trial.

Remark 1.14. The last equality in Equation 1.7 indicates that $\langle r \rangle$ is just the average number of spikes per trial divided by the trial duration.

Proposition 1.23. The average firing rate is equal to both the time average of $\mathrm{r(t)}$ and the trial average of the spike-count rate r, that is,

$$\langle r \rangle = \frac{1}{T}\int_0^T \mathrm{r(t)\,dt}. \tag{1.8}$$

Proof. By Definition 1.22,

$$\langle r \rangle = \frac{\langle n \rangle}{T} = \frac{1}{T}\int_0^T \langle \rho(\tau) \rangle\, d\tau = \frac{1}{T}\int_0^T \mathrm{r(t)\,dt},$$

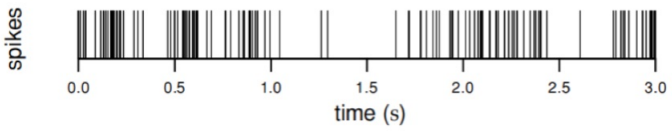
where the second equality follows from Proposition 1.16 and the third equality from Theorem 1.21. □

Remark 1.15. Whenever possible, we use the terms "firing rate", "spike-count rate", and "average firing rate" for $\mathrm{r}(t)$, r, and $\langle r \rangle$, respectively. In particular, we distinguish the spike-count rate r from the time-dependent firing rate $\mathrm{r}(t)$ by including the time argument in the latter expression (unless $\mathrm{r}(t)$ is independent of time).

1.2.2 Measuring Firing Rates

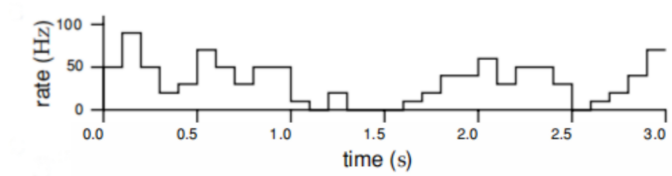
Notation 4. The firing rate $r(t)$ cannot be determined exactly from the limited data available from a finite number of trials. In addition, there is no unique way to approximate $r(t)$. We illustrate the methods of measuring firing rate by extracting firing rates from a single trial, but more accurate results could be obtained by averaging over multiple trials.

Example 1.24. We show an example of a 3 s spike train of the response of a neuron in the inferotemporal cortex recorded while a monkey watched a video. Neurons in the region of cortex where this recording was made are selective for complex visual images, including faces.



Algorithm 1.25. A simple algorithm of extracting an estimate of the firing rate from a spike train is to divide time into discrete bins of duration Δt, count the number of spikes within each bin, and divide by Δt.

Example 1.26. The figure in Example 1.26 shows the approximate firing rate computed using this procedure with a bin size of 100 ms. Note that with this procedure, the quantity being computed is really the spike-count firing rate over the duration of the bin, and that the firing rate $r(t)$ within a given bin is approximated by this spike-count rate. The binning and counting procedure illustrated in this figure generates an estimate of the firing rate that is a piecewise constant function of time, resembling a histogram.



Lemma 1.27. Because spike counts can take only integer values, the rates computed by this method will always be integer multiples of $1/\Delta t$, and thus they take discrete values. Decreasing the value of Δt increases temporal resolution by providing an estimate of the firing rate at more finely spaced intervals of time, but at the expense of decreasing the resolution for distinguishing different rates.

Algorithm 1.28. One algorithm to avoid quantized firing rates is to vary the bin size so that a fixed number of spikes appears in each bin. The firing rate is then approximated as that fixed number of spikes divided by the variable bin width.

Remark 1.16. Counting spikes in preassigned bins produces a firing-rate estimate that depends not only on the size of the time bins but also on their placement.

Algorithm 1.29. To avoid the arbitrariness in the placement of bins, the algorithm is to take a single bin or window of duration Δt and slide it along the spike train, counting the number of spikes within the window at each location which have a better temporal resolution.

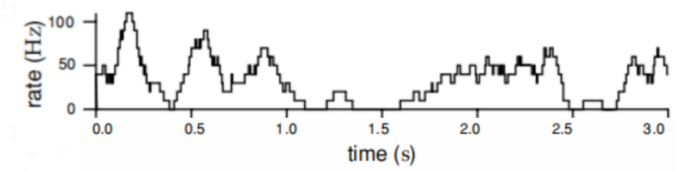
Example 1.30. The jagged curve in the figure of the Example 1.30 shows the result of sliding a 100 ms wide window along the spike train. The firing rate approximated in this way can be expressed as the sum of a window function over the times t_i for $i = 1, 2, \ldots, n$ when the n spikes in a particular sequence occurred,

$$r_{approx}(t) = \sum_{i=1}^{n} \omega(t - t_i), \tag{1.9}$$

where the window function is

$$\omega(t) = \begin{cases} 1/\Delta t & \text{if } -\Delta t/2 \le t \le \Delta t/2 \\ 0 & \text{otherwise.} \end{cases} \tag{1.10}$$

The jagged appearance of the curve is caused by the discontinuous shape of the window function used.



Proposition 1.31. The sum in Equation1.9 can also be written as the integral of the window function times the neural response function

$$r_{approx}(t) = \int_{-\infty}^{\infty} \omega(\tau)\rho(t - \tau)d\tau. \tag{1.11}$$

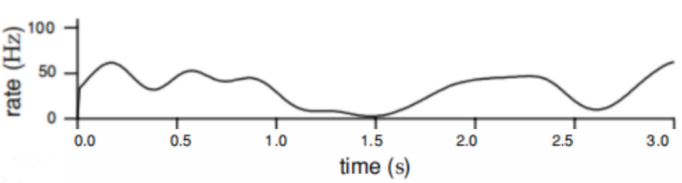
Proof. The proof directly follows from Equation 1.3. □

Definition 1.32. The integral in Equation 1.11 is called a *linear filter*, and the window function ω, also called the *filter kernel*, specifies how the neural response function evaluated at time $t - \tau$ contributes to the firing rate approximated at time t.

Example 1.33. Instead of the rectangular window function used in figure in Example 1.30, figure in Example 1.33 use a continuous window function like the Gaussian

$$\omega(\tau) = \frac{1}{\sqrt{2\pi}}\sigma_\omega \exp\left(-\frac{\tau^2}{2\sigma_\omega^2}\right), \tag{1.12}$$

which is used in Equation 1.9 generates a firing-rate estimate that is a smooth function of time. In this case, σ_ω controls the temporal resolution of the resulting rate, playing a role analogous to Δt.



Principle 1.34. A postsynaptic neuron monitoring the spike train of a presynaptic cell has access only to spikes that have previously occurred.

Definition 1.35. A window function or kernel is called *causal* when an approximation of the firing rate at time t that depends only on spikes fired before t can be calculated using a window function that vanishes when its argument is negative.

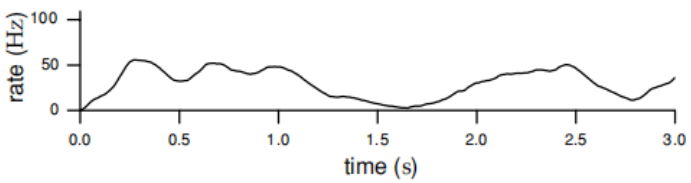
Definition 1.36. The *half-wave rectification* $[z]_+$ for any quantity z stands for,

$$[z]_+ = \begin{cases} z & \text{if} \quad z \geqslant 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1.13}$$

Example 1.37. One commonly used window function is the α function

$$w(\tau) = [\alpha^2 \tau \exp(-\alpha\tau)]_+ \tag{1.14}$$

where $1/\alpha$ determines the temporal resolution of the resulting firing-rate estimate. The figure below shows the firing rate approximated by such a causal scheme.



Remark 1.17. Note that the rate computed in Example 1.37 tends to peak later than the rate computed in Example 1.33 using a temporally symmetric window function.

1.2.3 Tuning Curve

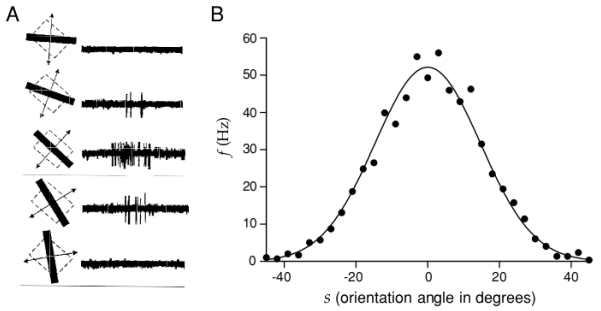
Remark 1.18. Neuronal responses typically depend on many different properties of a stimulus. In this chapter, we characterize responses of neurons as functions of just one of the stimulus attributes to which they may be sensitive. The value of this single attribute is denoted by s. In chapter 2, we consider more complete stimulus characterizations.

Definition 1.38. The *neural response tuning curve* is the average firing rate written as a function of s, $\langle r \rangle = f(s)$. The functional form of a tuning curve depends on the parameter s used to describe the stimulus.

Remark 1.19. A simple way of characterizing the response of a neuron is to count the number of action potentials fired during the presentation of a stimulus. This approach is most appropriate if the parameter s characterizing the stimulus is held constant over the trial. If we average the number of action potentials fired over (in theory, an infinite number of) trials and divide by the trial duration, we obtain the average firing rate, $\langle r \rangle$, defined in Equation 1.8.

Notation 5. Because tuning curves correspond to firing rates, they are measured in units of spikes per second or Hz.

Example 1.39. We show extracellular recordings of a neuron in the primary visual cortex (V1) of a monkey. While these recordings were being made, a bar of light was moved at different angles across the region of the visual field where the cell responded to light (see figure A). This region is called the receptive field of the neuron. Note that the number of action potentials fired depends on the angle of orientation of the bar.



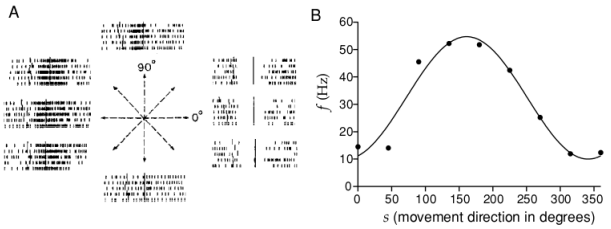
The dots in figure B indicate the average firing rate depends on the degrees of the orientation angle of the light bar stimulus. The data have been fitted by a response tuning curve of the form of *Gaussian tuning curve*

$$f(s) = r_{\max} \exp\left(-\frac{1}{2}\left(\frac{s - s_{\max}}{\sigma_f}\right)^2\right). \tag{1.15}$$

The curve in figure B is a fit using the Equation 1.15 with parameters $r_{\max} = 52.14$ Hz, $s_{\max} = 0°$, and $\sigma_f = 14.73°$, where s is the orientation angle of the light bar, $s_{\max}$ is the orientation angle evoking the maximum average response rate $r_{\max}$ (with $s - s_{\max}$ taken to lie in the range between $-90°$ and $+90°$), and σ_f determines the width of the tuning curve. The neuron responds most vigorously when a stimulus having $s = s_{\max}$ is presented, so we call $s_{\max}$ the preferred orientation angle of the neuron.

Remark 1.20. Tuning curves can also be measured for neurons in motor areas, in which case the average firing rate is expressed as a function of one or more parameters describing a motor action.

Example 1.40. We show an example of extracellular recordings from a neuron in primary motor cortex in a monkey that has been trained to reach in different directions. The stacked traces for each direction are rasters showing the results of five different trials. The horizontal axis in these traces represents time, and each mark indicates an action potential. The firing pattern of the cell, in particular the rate at which spikes are generated, is correlated with the direction of arm movement.encodes information about this aspect of the motor action.

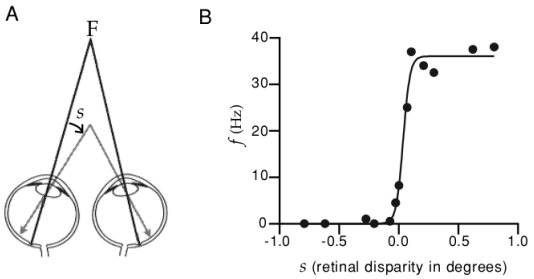


The dots in figure B indicate the average firing rate depends on the degrees of the direction of the arm movement. Here the data points have been fitted by a tuning curve in the form of a *cosine tuning curve*.

$$f(s) = [r_0 + (r_{\max} - r_0)\cos(s - s_{\max})]_+, \tag{1.16}$$

where s is the reaching angle of the arm, $s_{\max} = 161.25°$ is the reaching angle associated with the maximum response $r_{\max}$, and $r_0 = 32.34$ Hz is an offset or background firing rate that shifts the tuning curve up from the zero axis.

Example 1.41. We show a mode of retinal disparity. The gray lines with arrows show the location on each retina of an object located nearer than the fixation point F. The image from the fixation point falls at the fovea in each eye, the small pit where the black lines meet the retina. The image from a nearer object falls to the left of the fovea in the left eye and to the right of the fovea in the right eye. For objects farther away than the fixation point, this would be reversed. The disparity angle s is indicated in the figure.



The dots in figure B indicate average firing rate of a V1 neuron depends on retinal disparity and illustrates another important type of tuning curve. Here the data points have been fitted with a tuning curve called a *logistic function* or *sigmoidal function*,

$$f(s) = \frac{r_{\max}}{1 + \exp\left(\left(s_{1/2} - s\right)/\Delta_s\right)}. \tag{1.17}$$

In this case, s is the retinal disparity, the parameter $s_{1/2}$ is the disparity that produces a firing rate half as big as the maximum value $r_{\max}$, and Δ_s controls how quickly the firing rate increases as a function of s. If Δ_s is negative, the firing rate is a monotonically decreasing function of s rather than a monotonically increasing function.

1.2.4 Spike-Count Variability

Remark 1.21. Tuning curves allow us to predict the average firing rate, but they do not describe how the spike-count firing rate r varies about its mean value $\langle r \rangle = f(s)$ from trial to trial. While the map from stimulus to average response may be described deterministically, it is likely that single-trial responses such as spike-count rates can be modeled only in a probabilistic manner. Generally, r values can be generated from a probability distribution with mean $f(s)$.

Definition 1.42. The trial-to-trial deviation of r from $f(s)$ is considered to be noise, and such models are often called *noise models*.

Definition 1.43. The standard deviation for the noise distribution either can be independent of $f(s)$, in which case is called *additive noise*, or it can depend on $f(s)$. *Multiplicative noise* corresponds to having the standard deviation proportional to $f(s)$.

1.3 What Makes a Neuron Fire?

Remark 1.22. Response tuning curves characterize the average response of a neuron to a given stimulus. We now consider the complementary procedure of averaging the stimuli that produce a given response.

Remark 1.23. To average stimuli in this way, we need to specify what fixed response we will use to "trigger" the average. The most obvious choice is the firing of an action potential. Thus, we ask, "What, on average, did the stimulus do before an action potential was fired?" The resulting quantity, called the spike-triggered average stimulus, provides a useful way of characterizing neuronal selectivity.

1.3.1 Describing the Stimulus

Remark 1.24. Neurons responding to sensory stimuli face the difficult task of encoding parameters that can vary over an enormous dynamic range. To deal with such wide-ranging stimuli, sensory neurons often respond most strongly to rapid changes in stimulus properties and are relatively insensitive to steady-state levels. Steady-state responses are highly compressed functions of stimulus intensity, typically with logarithmic or weak power-law dependences. This compression has an interesting psychophysical correlate.

Notation 6. We use Δs denotes the "just noticeable" difference, which measures how different the intensity of two stimuli had to be for them to be reliably discriminated.

Principle 1.44 (Weber's law)**.** Δs is proportional to the magnitude of the stimulus s, so that $\Delta s/s$ is constant for a given stimulus.

Remark 1.25. Fechner suggested that noticeable differences set the scale for perceived stimulus intensities.

Principle 1.45 (Fechner's law)**.** Integrating Weber's law, the perceived intensity of a stimulus of absolute intensity s varies as $\log s$.

Remark 1.26. Sensory systems make numerous adaptations, using a variety of mechanisms, to adjust to the average level of stimulus intensity. When a stimulus generates such adaptation, the relationship between stimulus and response is often studied in a potentially simpler regime by describing responses to fluctuations about a mean stimulus level.

Assumption 1.46. We frequently impose this condition that $s(t)$ satisfies

$$\frac{1}{T}\int_0^T s(t)dt = 0.$$

Remark 1.27. Our analysis of neural encoding involves two different types of averages: averages over repeated trials that employ the same stimulus, which we denote by angle brackets, and averages over different stimuli. We could introduce a second notation for averages over stimuli, but this can be avoided when using time-dependent stimuli.

Notation 7. Instead of presenting a number of different stimuli and averaging over them, we string together all of the stimuli we wish to consider into a single time-dependent stimulus sequence and average over time. Thus, *stimulus averages* are replaced by *time averages*.

Assumption 1.47 (Periodic Stimulus)**.** In the following, we assume that the stimulus are time-translationally invariant, that is, $s(T+\tau) = s(\tau)$ for any τ.

Proposition 1.48. If integrals involving the stimulus are time-translationally invariant, then for any function h and time interval τ

$$\int_0^T h(s(t+\tau))dt = \int_0^T h(s(t))dt. \tag{1.18}$$

Proof.

$$\int_0^T h(s(t+\tau))dt = \int_\tau^{T+\tau} h(s(t))dt = \int_0^T h(s(t))dt,$$

where the last follows from the stimulus periodicity. □

1.3.2 The Spike-Triggered Average

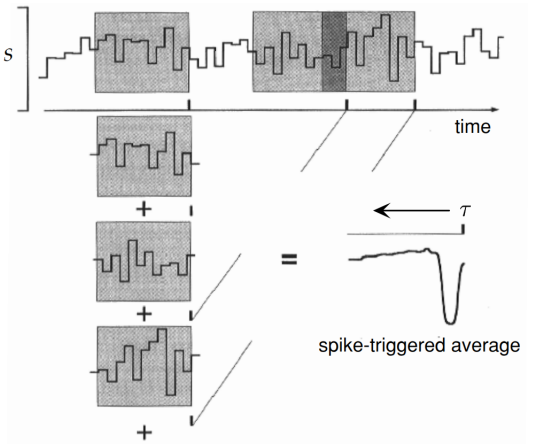
Definition 1.49. The *spike-triggered average stimulus* (or *spike-triggered average*) $C(\tau)$ is the average value of the stimulus s a time interval τ before a spike is fired.

Proposition 1.50. The spike-triggered average $C(\tau)$ satisfies

$$C(\tau) = \left\langle \frac{1}{n}\sum_{i=1}^{n} s(t_i - \tau) \right\rangle \approx \frac{1}{\langle n \rangle} \left\langle \sum_{i=1}^{n} s(t_i - \tau) \right\rangle, \tag{1.19}$$

where n is the total number of spikes on each trial and t_i is the time when the spike occurrs for $i = 1, 2, \cdots, n$. If n is large, it is well approximated by $\langle n \rangle$.

Example 1.51. The following figure provides a schematic description of the computation of the spike-triggered average. Each time a spike appears, the stimulus in a time window preceding the spike is recorded. Although the range of τ values in Equation 1.19 is unlimited, the response is typically affected only by the stimulus in a window a few hundred milliseconds wide immediately preceding a spike. In practice, the stimulus is recorded only over a finite time period, as indicated by the shaded areas in the figure below. The recorded stimuli for all spikes are then summed and the procedure is repeated over multiple trials to produce the waveform shown at the lower right, which is the average stimulus before a spike.



Assumption 1.52. We expect $C(\tau)$ to approach 0 for positive τ values larger than the correlation time between the stimulus and the response. If the stimulus has no temporal correlations with itself, we also expect $C(\tau)$ to be 0 for $\tau < 0$, because the response of a neuron cannot depend on future stimuli.

Remark 1.28. We make use of this approximation in Equation 1.19 because it allows us to relate the spike-triggered average to other quantities commonly used to characterize the relationship between stimulus and response (see Proposition 1.53).

Proposition 1.53. Using the approximate expression for $C(\tau)$ in Equation 1.19, we find

$$C(\tau) = \frac{1}{\langle n \rangle} \int_0^T \mathrm{r}(t) s(t-\tau) dt. \tag{1.20}$$

Proof. By Theorem 1.15, the spike-triggered average stimulus can be expressed as an integral of the stimulus times the neural response function $\rho(t)$. Thus,

$$C(\tau) = \frac{1}{\langle n \rangle} \int_0^T \langle \rho(t) \rangle s(t-\tau) dt = \frac{1}{\langle n \rangle} \int_0^T \mathrm{r}(t) s(t-\tau) dt,$$

where the second step follows from the equivalence of $\langle \rho(t) \rangle$ and $\mathrm{r}(t)$ within integrals. □

Remark 1.29. Equation 1.20 allows us to relate the spike-triggered average to the correlation function of the firing rate and the stimulus.

Definition 1.54. The *correlation function* of the any two well-behaved functions $f(t)$ and $g(t)$ on $[0, T]$ is

$$Q_{fg}(\tau) = \frac{1}{T} \int_0^T f(t) g(t+\tau) dt. \tag{1.21}$$

Remark 1.30. Correlation functions are a useful way of determining how two quantities that vary over time are related to one another. The two quantities being related are evaluated at different times, one at time t and the other at time $t + \tau$. The correlation function is then obtained by averaging their product over all t values, and it is a function of τ.

Proposition 1.55. The correlation function of the firing rate and the stimulus (called *firing-rate stimulus correlation function*) is

$$Q_{\mathrm{rs}} = \frac{1}{T}\int_0^T \mathrm{r}(t)s(t+\tau)dt. \tag{1.22}$$

Proposition 1.56. $C(\tau)$ satisfies

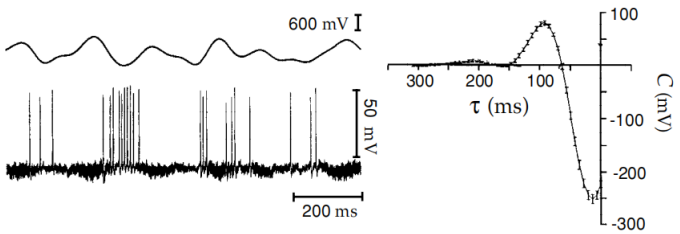
$$C(\tau) = \frac{1}{\langle r\rangle}Q_{\mathrm{rs}}(-\tau), \tag{1.23}$$

where $\langle r\rangle = \langle n\rangle/T$ is the average firing rate over the set of trials.

Proof. This can be obtained by comparing Equations 1.20 and 1.22. □

Notation 8. Because the argument of the correlation function in Equation 1.23 is $-\tau$, the spike-triggered average stimulus is often called the *reverse correlation function*.

Example 1.57. The following figure shows the spike-triggered average stimulus for a neuron in the electrosensory lateral-line lobe of the weakly electric fish *Eigenmannia*. Fluctuating electrical potentials, such as that shown in the upper left trace of the figure below, elicit responses from electrosensory lateral-line lobe neurons, as seen in the lower left trace. The spike-triggered average stimulus, plotted at the right, indicates that, on average, the electric potential made a positive upswing followed by a large negative deviation prior to a spike being fired by this neuron.



Remark 1.31. The spike-triggered average stimulus is widely used to study and characterize neural responses. Because $C(\tau)$ is the average value of the stimulus at a time τ before a spike, larger values of τ represent times farther in the past relative to the time of the triggering spike. For this reason, we spike-triggered averages with the time axis going backward compared to the normal convention. This allows the average spike-triggering stimulus to be read off from the plots in the usual left-to-right order.

Remark 1.32. The results obtained by spike-triggered averaging depend on the particular set of stimuli used during an experiment. How should this set be chosen? In chapter 2, we show that there are certain advantages to using a stimulus that is uncorrelated from one time to the next, a white-noise stimulus. A heuristic argument supporting the use of such stimuli is that in asking what makes a neuron fire, we may want to sample its responses to stimulus fluctuations at all frequencies with equal weight (i.e., equal power), and this is one of the properties of white-noise stimuli.

1.3.3 White-Noise Stimuli

Definition 1.58. The defining characteristic of *white-noise stimulus* is that its value at any one time is uncorrelated with its value at any other time.

Proposition 1.59. The stimulus-stimulus correlation function (also called the *stimulus autocorrelation*) for white-noise stimulus $s(t)$ can be expressed by

$$Q_{ss}(\tau) = \sigma_s^2\delta(\tau) \tag{1.24}$$

with some constant σ_s.

Proof. By Definition 1.54, we have

$$Q_{ss}(\tau) = \frac{1}{T}\int_0^T s(t)s(t+\tau)dt. \tag{1.25}$$

Just as a correlation function provides information about the temporal relationship between two quantities, so an autocorrelation function tells us about how a quantity at one time is related to itself evaluated at another time. For white noise, the stimulus autocorrelation function is 0 in the range $-T/2 < \tau < T/2$ except when $\tau = 0$, thus, over this range we have Equation 1.24. □

Definition 1.60. The power in a signal as a function of its frequency is called the *power spectrum* or *power spectral density*.

Remark 1.33. White noise has a flat power spectrum.

Definition 1.61. The *power spectrum* for a stimulus $s(t)$ is the Fourier transform of the autocorrelation function of $s(t)$.

$$\widetilde{Q}_{ss}(\omega) = \frac{1}{T}\int_{-T/2}^{T/2} Q_{ss}(\tau)\exp(i\omega\tau)d\tau. \tag{1.26}$$

Remark 1.34. Because we have defined the stimulus as periodic outside the range of the trial T, we have used a finite-time Fourier transform and ω should be restricted to values that are integer multiples of $2\pi/T$.

Lemma 1.62. The power spectrum for a white-noise stimulus $s(t)$ is

$$\widetilde{Q}_{ss}(\omega) = \frac{\sigma_s^2}{T}, \tag{1.27}$$

which is the defining characteristic of white noise; its power spectrum is independent of frequency.

Proof. Using the fact that $Q_{ss}(\tau) = \sigma_s^2\delta(\tau)$ for white noise, we have

$$\widetilde{Q}_{ss}(\omega) = \frac{\sigma_s^2}{T}\int_{-T/2}^{T/2}\delta(t)\exp(i\omega\tau)d\tau = \frac{\sigma_s^2}{T}.$$

□

Proposition 1.63. Equation 1.24 is equivalent to the statement that white noise has equal power at all frequencies.

Solution. This conclusion is directly derived from Lemma 1.62.

Proposition 1.64. The *power spectrum* for a stimulus $s(t)$ satisfies

$$\widetilde{Q}_{ss}(\omega) = |\widetilde{s}(\omega)|^2. \tag{1.28}$$

Proof. Using the definition of the stimulus autocorrelation function, we can also write

$$\begin{aligned}
&\widetilde{Q}_{ss}(\omega)\\
=&\frac{1}{T}\int_0^T s(t)\frac{1}{T}\int_{-T/2}^{T/2} s(t+\tau)e^{i\omega\tau}d\tau dt\\
=&\frac{1}{T}\int_0^T s(t)e^{-i\omega t}\frac{1}{T}\int_{-T/2+t}^{T/2+t} s(t+\tau)e^{i\omega(t+\tau)}d(t+\tau)dt\\
=&\frac{1}{T}\int_0^T s(t)e^{-i\omega t}\frac{1}{T}\int_{-T/2}^{T/2} s(\tau)e^{i\omega(\tau)}d(\tau)dt\\
=&\frac{1}{T}\int_0^T s(t)e^{-i\omega t}dt\frac{1}{T}\int_{-T/2}^{T/2} s(\tau)e^{i\omega(\tau)}d(\tau),
\end{aligned}$$

where the seond step and third step follow from the variable substitution and the periodicity of the stimulus. The first integral on the right side of the forth equality is the complex conjugate of the Fourier transform of the stimulus,

$$\widetilde{s}(\omega) = \frac{1}{T}\int_0^T s(t)\exp(i\omega\tau)d\tau. \tag{1.29}$$

The second integral, because of the periodicity of the integrand (when ω is an integer multiple of $2\pi/T$) is equal to $\widetilde{s}(\omega)$. Therefore,

$$\widetilde{Q}_{ss}(\omega) = |\widetilde{s}(\omega)|^2, \tag{1.30}$$

which provides another definition of the stimulus power spectrum. It is the absolute square of the Fourier transform of the stimulus. □

Remark 1.35. No physical system can generate noise that is white to arbitrarily high frequencies. Approximations of white noise that are missing high-frequency components can be used, provided the missing frequencies are well above the sensitivity of the neuron under investigation.

Notation 9. To approximate white noise, we consider times that are integer multiples of a basic unit of duration Δt, that is, times $t = m\Delta t$ for $m = 1, 2, \cdots, M$ where $M\Delta t = T$. The function $s(t)$ is then constructed as a discrete sequence of stimulus values.

Proposition 1.65. In terms of the discrete-time values $s(t) = s_m$ for $(m-1)\Delta t \le t < m\Delta t$, the condition that the stimulus is uncorrelated is

$$\frac{1}{M}\sum_{m=1}^{M} s_m s_{m+p} = \begin{cases}\sigma_s^2/\Delta t & \text{if} \quad p = 0\\ 0 & \text{otherwise.}\end{cases} \tag{1.31}$$

Remark 1.36. The factor of $1/\Delta t$ on the right side of Equation 1.31 reproduces the δ function of Equation 1.24 in the limit $\Delta t \to 0$. For approximate white noise, the autocorrelation function is 0 except for a region around $\tau = 0$ with width of order Δt.

Exercise 1.66. The binning of time into discrete intervals of size Δt means that the noise generated has a flat power spectrum only up to frequencies of order $1/(2\Delta t)$.

Algorithm 1.67. An approximation to white noise can be generated by choosing each s_m independently from a probability distribution with mean 0 and variance $\sigma_s^2/\Delta t$.

Remark 1.37. Any reasonable probability function satisfying two conditions in Rule 1.67 can be used to generate the stimulus values within each time bin. The factor of $1/\Delta t$ in the variance indicates that the variability must be increased as the time bins get smaller.

Example 1.68. A special class of white-noise stimuli, Gaussian white noise, results when the probability distribution used to generate the s_m values is a Gaussian function.

Remark 1.38. Although Equations 1.26 and 1.30 are both sound, they do not provide a statistically efficient method of estimating the power spectrum of discrete approximations to white-noise sequences generated by the methods described in this chapter.

Algorithm 1.69. The apparently natural procedure of taking a white-noise sequence $s(m\Delta t)$ for $m = 1, 2, \cdots, T/\Delta t$, and computing the square amplitude of its Fourier transform at frequency ω, is

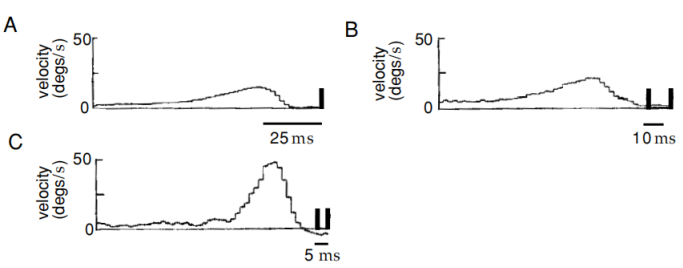
$$\frac{\Delta t}{T}\left|\sum_{m=1}^{T/\Delta t} s(t)\exp(-i\omega m\Delta t)\right|^2.$$

Remark 1.39. The procedure in Algorithm 1.69 is a biased and extremely noisy way of estimating $\widetilde{Q}_{ss}(\omega)$. This estimator is called the *periodogram*. The statistical problems with the periodogram, and some of the many suggested solutions, are discussed in almost any textbook on spectral analysis.

1.3.4 Multiple-Spike-Triggered Averages and Spike-Triggered Correlations

Remark 1.40. In addition to triggering on single spikes, stimulus averages can be computed by triggering on various combinations of spikes.

Example 1.70. The following pictures show Single- and multiple-spike-triggered average stimuli for a blowfly H1 neuron responding to a moving visual image. Here the plot A is the average stimulus velocity triggered on a single spike. The plot B is the average stimulus velocity before two spikes with a separation of 10 ± 1 ms. Plot C is the average stimulus before two spikes with a separation of 5 ± 1 ms.



In the case of B, the two-spike average is similar to the sum of two single-spike-triggered average stimuli displaced from one another by 10 ms. Thus, for 10 ms separations, two spikes occurring together tell us no more as a two-spike unit than they would individually. This result changes when shorter separations are considered. In the case of C, the average stimulus triggered on a pair of spikes separated by 5 ms is not the same as the sum of the average stimuli for each spike separately.

Remark 1.41. Spike-triggered averages of other stimulus-dependent quantities can provide additional insight into neural encoding, for example, spike-triggered average autocorrelation functions. Obviously, spike-triggered averages of higher-order stimulus combinations can be considered as well.

1.4 Spike-Train Statistics

Remark 1.42. A complete description of the stochastic relationship between a stimulus and a response would require us to know the probabilities corresponding to every sequence of spikes that can be evoked by the stimulus.

Lemma 1.71. The probability that z takes a value between z and $z + \Delta z$, for small Δz (strictly speaking, as $\Delta z \to 0$), is equal to $p[z]\Delta z$, where $p[z]$ is called a probability density.

Notation 10. Throughout this book, we use the notation $P[\ ]$ to denote probabilities and $p[\ ]$ to denote probability densities.

Theorem 1.72. The probability of a spike sequence appearing is proportional to the probability density of spike times, $p[t_1, t_2, ..., t_n]$. More precisely, the probability $P[t_1, t_2, ..., t_n]$ that a sequence of n spikes occurs with spike i falling between times t_i and $t_i + \Delta t$ for $i = 1, 2, \ldots, n$ is given in terms of this density by the relation

$$P[t_1, t_2, ..., t_n] = p[t_1, t_2, ..., t_n](\Delta t)^n. \tag{1.32}$$

Proof. By the relationship between probability and probability density, we have

$$\begin{aligned} P[t_1, t_2, ..., t_n] &= \int\!\!\cdots\!\!\int_{\substack{t_i - \frac{\Delta t}{2} \le s_i \le t_i - \frac{\Delta t}{2} \\ 1 \le i \le n}} p[s_1, s_2, ..., s_n] ds_1 ... ds_{n-1} ds_n \\ &= p[t_1, t_2, ..., t_n](\Delta t)^n, \end{aligned}$$

where the second equality follows from the integral mean value theorem (as $\Delta t \to 0$). □

Definition 1.73. A stochastic process that generates a sequence of events, such as action potentials ,is called a *point process.*

Remark 1.43. In general, the probability of an event occurring at any given time could depend on the entire history of preceding events.

Definition 1.74. If the dependence on history events extends only to the immediately preceding event, so that the intervals between successive events are independent, the point process is called a *renewal process.*

Definition 1.75. If there is no dependence renewal process at all on preceding events, so that the events themselves are statistically independent, the point process is called a *Poisson process.* The Poisson process is *homogeneous* if the firing rate is constant over time, and is *inhomogeneous* if it involves a time-dependent firing rate.

1.4.1 The Homogeneous Poisson Process

Notation 11. Denote the firing rate for a homogeneous Poisson process by $r(t) = r$, because it is independent of time.

Notation 12. Denote the probality that an arbitrary sequence of exactly n spikes occurs within a trial of duration T by $P_T[n]$.

Theorem 1.76. For a homogeneous Poisson process, the distribution of the spike count within a trial of duration T is a Possion distribution,

$$P_T[n] = \frac{(rn)^n}{n} \exp(-rT). \tag{1.33}$$

Proof. To compute $P_T[n]$, we divide the time T into M bins of size $\Delta t = T/M$. We assume that Δt is small enough so that we never get two spikes within any one bin because, at the end of the calculation,we take the limit $\Delta t \to 0$. $P_T[n]$ is the product of three factors: the probability of generating n spikes within a specified set of the M bins, $\frac{M!}{(M-n)!n!}$; the probability of not generating spikes in the remaining $M - n$ bins, $(r\Delta t)^n$; a combinatorial factor equal to the number of ways of putting n spikes into M bins, $(1 - r\Delta t)^{M-n}$. That is,

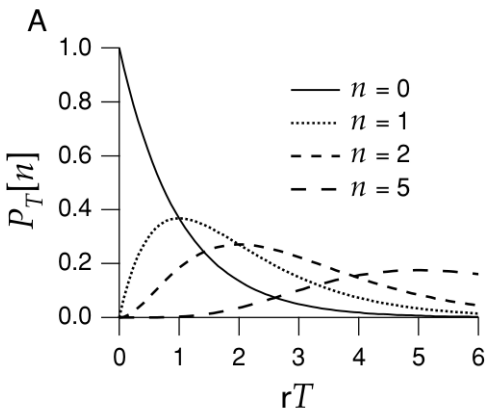
$$P_T[n] = \lim_{\Delta t \to 0} \frac{M!}{(M-n)!n!} (r\Delta t)^n (1 - r\Delta t)^{M-n}. \tag{1.34}$$

As $\Delta t \to 0, M$ grows without bound because $M\Delta t = T$. Because n is fixed, we can write $M - n \approx M = T/\Delta t$. Using this approximation and defining $\epsilon = -r\Delta t$, we find that

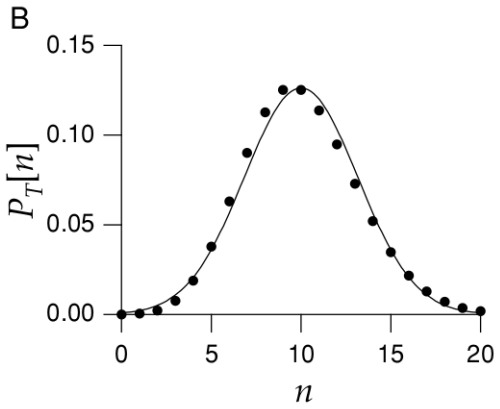
$$\lim_{\Delta t \to 0} (1 - r\Delta t)^{M-n} = \lim_{\epsilon \to 0} (((1 + \epsilon)^{\frac{1}{\epsilon}})^{-rT} = \exp(-rT). \tag{1.35}$$

For large M, $\frac{M!}{(M-n)!} \approx M^n = (T/\Delta t)^n$ completes the proof. □

Example 1.77. The probabilities $P_T[n]$, for a few n values, are plotted as a function of rT in the following firgue. Note that as n increase, the probability reaches its maximum at larger T values and that large n values are more likely than small ones for large T.



Example 1.78. The following figure shows the probabilities of various numbers of spikes occurring when the average number of spikes is 10. For large rT, which corresponds to a large expected number of spikes, the Poisson distribution approaches a Gaussian distribution with mean and variance equal to rT. This figure shows that this approximation is already quite good for $rT = 10$.



Theorem 1.79. The probability $P[t_1, t_2, ..., t_n]$ can be expressed in terms of another probability function $P_T[n]$. Assuming that the spike times are ordered $0 \leq t_1 \leq t_2 \leq ... \leq t_n \leq T$, the relationship is

$$P[t_1, t_2, ..., t_n] = n! P_T[n] \left(\frac{\Delta t}{T}\right)^n. \tag{1.36}$$

Proof. The probability of docking in a specific time order $(t_1, t_2, ..., t_n)$ is $\left(n(\frac{\Delta t}{T})(n-1)(\frac{\Delta t}{T})\dots 1(\frac{\Delta t}{T})\right)$. Thus,

$$\begin{aligned} P[t_1, t_2, ..., t_n] &= P_T[n](n(\frac{\Delta t}{T})(n-1)(\frac{\Delta t}{T})\dots 1(\frac{\Delta t}{T})) \\ &= n! P_T[n] \left(\frac{\Delta t}{T}\right)^n, \end{aligned}$$

which completes the proof. □

Corollary 1.80. For spikes counted over an interval of duration T, the variance of the spike count is

$$\sigma_n^2 = \langle n^2 \rangle - \langle n \rangle^2 = rT. \tag{1.37}$$

Proof. The average number of spikes generated by a Poisson process with constasnt rate r over a time T is

$$\langle n \rangle = \sum_{n=0}^{\infty} n P_T[n] = \sum_{n=0}^{\infty} \frac{n(rT)^n}{n!} \exp(-rT). \tag{1.38}$$

and the variance in the spike count is

$$\sigma_n^2(T) = \sum_{n=0}^{\infty} n^2 P_T[n] - \langle n \rangle^2 = \sum_{n=0}^{\infty} \frac{n^2(rT)^n}{n!} \exp(-rT) - \langle n \rangle^2. \tag{1.39}$$

To compute the quantities, we need to calculate the two sums appearing in these Equations. A good way to do this is to compute the moment-generating function

$$g(\alpha) = \sum_{n=0}^{\infty} \frac{(rT)^n \exp(\alpha n)}{n!} \exp(-rT). \tag{1.40}$$

The kth derivative of g with respect to α, evaluated at the point $\alpha = 0$, is

$$\left.\frac{dg}{d\alpha^k}\right|_{\alpha=0} = \sum_{n=0}^{\infty} \frac{n^k(rT)^n}{n!} \exp(-rT), \tag{1.41}$$

so once we have computed g, we need to calculate only its first and second derivative to determine the sums we need. Rearranging the terms a bit, and recalling that $\exp(z) = \sum z^n/n!$, we find

$$g(\alpha) = \exp(-rT) \sum_{n=0}^{\infty} \frac{(rT\exp(\alpha))^n}{n!} = \exp(-rT)\exp(rTe^{\alpha}). \tag{1.42}$$

The derivatives are then

$$\frac{dg}{d\alpha} = rTe^{\alpha} \exp(-rT) \exp(rTe^{\alpha}) \tag{1.43}$$

and

$$\frac{dg}{d\alpha^2} = (rTe^{\alpha})^2 \exp(-rT)\exp(rTe^{\alpha}) + rTe^{\alpha}\exp(-rT)\exp(rTe^{\alpha}). \tag{1.44}$$

Evaluating these at $\alpha = 0$ and putting the results into Equation 1.38 and 1.39 gives the result $\langle n \rangle = rT$ and $\sigma_n^2(T) = (rT)^2 + rT - (rT)^2 = rT$. □

Definition 1.81. The ratio of the variance and mean of the spike count, $\sigma_n^2/\langle n \rangle$, is called the *Fano factor*.

Example 1.82. The Fano factor takes the value 1 for a homogeneous Poisson process, independent of the time interval T.

Lemma 1.83. The interspike interval distribution for a homogeneous Poisson spike train is an exponential. Equivalently, the probability of an interspike interval falling between τ and $\tau + \Delta t$ is

$$P[\tau \leq t_{i+1} - t_i < \tau + \Delta t] = r\Delta t \ \exp(-r\tau). \tag{1.45}$$

Proof. Suppose that a spike occurs at a time t_i for some value of i. The probability of a homogeneous Poisson process generating the next spike somewhere in the interval

$$t_i + \tau \leq t_{i+1} \leq t_i + \tau + \Delta t,$$

for small Δt, is the probabilities that no spike is fired for a time τ, times the probability, $r\Delta t$, of generating a spike within the following small interval Δt. From Equation 1.33, with $n = 0$, the probability of not firing a spike for period τ is $\exp(-r\tau)$. So the probability of an interspike interval falling between τ and $\tau + \Delta t$ is $r\Delta t\ \exp(-r\tau)$. This completes the proof. □

Theorem 1.84. The mean interspike interval and the variance of the interspike intervals satisfy

$$\langle \tau \rangle = \frac{1}{r}, \tag{1.46}$$

and

$$\sigma_\tau^2 = \frac{1}{r^2}. \tag{1.47}$$

Proof. By the interspike interval distribution of a homogeneous Poisson spike train, we have

$$\langle \tau \rangle = \int_0^\infty \tau r\ \exp(-r\tau) d\tau \overset{s=r\tau}{=} \frac{1}{r} \int_0^{infty} s e^{-s} ds = \frac{1}{r},$$

where the third equality follows from the integration by parts, and

$$\begin{aligned} &\int_0^\infty \tau^2 r\ \exp(-r\tau) d\tau \overset{s=r\tau}{=} \frac{1}{r^2} \int_0^\infty s^2 e^{-s} ds \\ &= \frac{1}{r^2} \left(-s^2 e^s \big|_0^\infty + \int_0^\infty 2s e^{-s} ds \right) = \frac{2}{r^2}, \end{aligned}$$

where the second equality follows from the integration by parts. Thus,

$$\sigma_\tau^2 = \int_0^\infty \tau^2 r\ \exp(-r\tau) d\tau - \langle \tau \rangle^2 = \frac{2}{r^2} - \frac{1}{r^2}.$$

This completes the proof. □

Definition 1.85. The ratio of the standard deviation and the mean of interspike interval distribution

$$C_V = \frac{\sigma_\tau}{\langle \tau \rangle} \tag{1.48}$$

is called the *coefficient of variation*

Remark 1.44. The coefficient of variation takes the value 1 for a homogeneous Poisson process. This is a necessary, though not sufficient, condition to identify a Poisson spike train. Recall that the Fano factor for a Poisson process is also 1.

Exercise 1.86. For any renewal process, the Fano factor evaluated over long time intervals approaches the value C_V^2.

1.4.2 The Spike-Train Autocorrelation Funciton

Definition 1.87. The *spike-train autocorrelation function*,

$$Q_{\rho\rho}(\tau) = \frac{1}{T} \int_0^T \langle (\rho(t) - \langle r \rangle)(\rho(t+\tau) - \langle r \rangle) \rangle dt, \tag{1.49}$$

is the autocorrelation of the neural response function of Equation 1.1 with its average over time and trials substracted out.

Theorem 1.88. The autocorrelation function for a Poisson spike train generated at a constant rate $\langle r \rangle = r$ is

$$Q_{\rho\rho}(\tau) = r\delta(\tau). \tag{1.50}$$

Proof. The spike-train autocorrelation function is constructed from data in the form of a histogram by dividing time into bins. The value of the histogram for a bin labeled with a positive or negative integer m is computed by determining the number of the times that any two spikes in the train are separated by a time interval lying between $(m-1/2)\Delta t$ and $(m+1/2)\Delta$ with Δt the bin size. This includes all pairings, even between a spike and itself. We call this number N_m. If the intervals between the n^2 spike pairs in the train were uniformly distributed over the range from 0 to T, there would be $n^2\Delta t/T$ intervals in each bin. This uniform term is removed from the autocorrelation histogram by subtracting $n^2\Delta t/T$ from N_m for all m. The spike-train autocorrelation histogram is then defined by dividing the resulting numbers by T, so the value of the histogram in bin m is $H_m = N_m/T - n^2\Delta/T^2$. For small bin sizes, the $m = 0$ term in the histogram counts the average number of spikes, that is $N_m = \langle n \rangle$ and in the limit $\Delta t \to 0$, $H_0 = \langle n \rangle / T$ is the average firing rate $\langle r \rangle$. Because other bins have H_m of order Δt, large $m = 0$ term is often removed from histogram plots. The spike-train autocorrlation function is defined as $H_m/\Delta t$ in the limit $\Delta t \to 0$, and it has the units of a firing rate squared. In this limit, the $m = 0$ bin becomes a δfuncitn, $H_0/\Delta t \to \langle r \rangle \delta(\tau)$.
As we can have seen, the distribution of interspikde intervals for adjacent spikes in a homogeneous Poisson spike train is exponential(Equation 1.45). By contrast, the intervals between any two spikes(not necessarily adjacent) in such a train are uniformly distributed. As a result, the subtraction procedure outlined above gives $H_m = 0$ for all bins except for the $m = 0$ bin that contains the contribution of the zero intervals between spikes and themselves. The autocorrlation function for a Poisson spike train generated at a constant rate $\langle r \rangle = r$ is $Q_{\rho\rho}(\tau) = r\delta(\tau)$. □

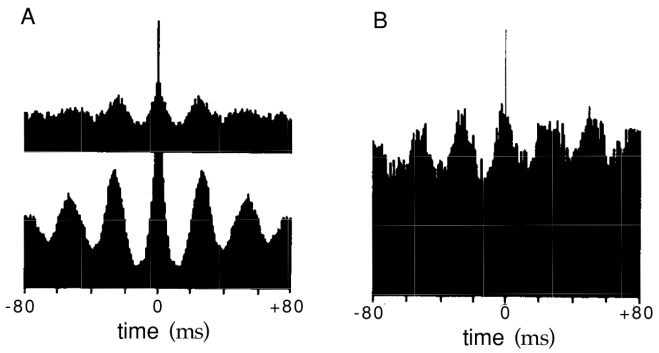
Definition 1.89. The *spike-train cross-correlation function* between spike trains from two different neurons can be defined by,

$$Q_{\rho_1\rho_2}(\tau) = \frac{1}{T} \int_0^T \langle (\rho_1(t) - \langle r_1 \rangle)(\rho_2(t+\tau) - \langle r_2 \rangle) \rangle dt, \tag{1.51}$$

where $\rho_1(t)$ and $\rho_2(t)$ are neural response functions of these two neurons, and $\langle r_1 \rangle$ and $\langle r_2 \rangle$ are their average firing rates.

Remark 1.45. The spike-train autocorrelation function is an even function of τ, $Q_{\rho\rho}(\tau) = Q_{\rho\rho}(-\tau)$, but the cross-correlation function is not necessarily even.

Example 1.90. The following figures show autocorrelation and cross-correlation histograms for neurons in the primary visual cortex of a cat. Here the plot A is autocorrelation histograms for neurons recorded in the right (upper) and left (lower) hemispheres show a periodic pattern indicating oscillations at about 40 Hz. The lower diagram indicates stronger oscillations in the left hemisphere. The plot B is the cross-correlation histogram for these two neurons shows that their oscillations are synchronized with little time delay.



Remark 1.46. A peak at zero interval in a cross-correlation function signifies that the two neurons are firing synchronously. Asymmetric shifts in this peak away from 0 result from fixed delays between the firing of the two neurons, and they indicate *nonsynchronous but phase-locked firing.* Periodic structure in either an autocorrelation or a cross-correlation function or histogram indicates that the *firing probability oscillates.*

1.4.3 The Inhomogeneous Poisson Process

Theorem 1.91. The probability density of the inhomogeneous Poisson Process for n spike times is

$$p[t_1, t_2, ..., t_n] = \exp\left(-\int_0^T r(t)dt\right)\prod_{i=1}^{n} r(t_i), \qquad (1.52)$$

where the spike times are ordered $0 \le t_1 \le t_2 \le ... \le t_n \le T$.

Proof. The probability density for a particular spike sequence with spike times t_i for $i = 1, 2, ..., n$ is obtained from the corresponding probability distribution by multiplying the probability that the spikes occur when they do by the probability that no other spikes occur.We begin by computing the probability that no spikes are generated during the time interval from t_i to t_{i+1} between two adjacent spikes. We determine this by dividing the interval into M bins of size Δt and setting $M\Delta t = t_{i+1} - t_i$. We will ultimately take the limit $\Delta t \to 0$. The firing rate during bin m within this interval is $r(t_i + m\Delta t)$. Because the probability of firing a spike in this bin is $r(t_i+m\Delta t)\Delta t$, the probabilities of not firing a spike is $1 - r(t_i + m\Delta t)\Delta t$. To have no spikes during the entire interval, we must string together M such bins, and the probability of this occurring is the product of the individual probabilities,

$$P[\text{no spikes}] = \prod_{m=1}^{M}(1 - r(t_i + m\Delta t)\Delta t). \qquad (1.53)$$

We evaluate this expression by taking its logarithm,

$$\ln P[\text{no spikes}] = \sum_{m=1}^{M}\ln(1 - r(t_i + m\Delta t)\Delta t), \qquad (1.54)$$

using the fact that the logarithm of a product is the sum of the logarithms of the multiplied terms. Using the approximation $\ln(1 - r(t_i + m\Delta t)\Delta t) \approx -r(t_i + m\Delta t)\Delta t$, valid for small Δt, we can simplify this to

$$\ln P[\text{no spikes}] = -\sum_{m=1}^{M} r(t_i + m\Delta t)\Delta t. \qquad (1.55)$$

In the limit $\Delta t \to 0$, the approximation becomes exact and this sum becomes the integral of $r(t)$ from t_i to t_{i+1},

$$\ln P[\text{no spikes}] = -\int_{t_i}^{t_{i+1}} r(t)dt. \qquad (1.56)$$

Exponentiating this Equation gives the result we need,

$$P[\text{no spikes}] = \exp\left(-\int_{t_i}^{t_{i+1}} r(t)dt\right). \qquad (1.57)$$

The probability density $p[t_1, t_2, ..., t_n]$ is the product of the densities for the individual spikes and the probabilities of not generating spikes during the interspikde intervals, between time 0 and the first spike, and between the time of the last spike and the end of the trial period:

$$p[t_1, t_2, ...t_n] = \exp\left(-\int_0^{t_1} r(t)dt\right)\exp\left(-\int_{t_n}^{T} r(t)dt\right)\times r(t_n)\prod_{i=1}^{n-1} r(t_i)\exp\left(-\int_{t_i}^{t_{i+1}} r(t)dt\right). \qquad (1.58)$$

The exponentials in this expression all combine because the product of exponentials is the exponential of the sum, so the different integrals in this sum add up to form a single integral:

$$\begin{aligned}&\exp\left(-\int_0^{t_1} r(t)dt)\right)\exp\left(-\int_{t_n}^{T} r(t)dt\right)\prod_{i=1}^{n-1}\exp\left(-\int_{t_i}^{t_{i+1}} r(t)dt\right)\\ &= \exp\left(-\left(\int_0^{t_1} r(t)dt + \sum_{i=1}^{n-1}\int_{t_i}^{t_{i+1}} r(t)dt + \int_{t_n}^{T} r(t)dt\right)\right)\\ &= \exp\left(-\int_0^{T} r(t)dt\right).\end{aligned} \qquad (1.59)$$

Substituting this into Equation 1.58 gives the result in Equation 1.52. □

Remark 1.47. The eqution 1.36 is a special case of Equation 1.52.

1.4.4 The Poisson Spike Generator

Rule 1.92. Spike sequences can be simulated by using some estimate of the firing rate, $r_{\text{est}}(t)$, predicted from knowledge of the stimulus, to drive a Poisson process.

Algorithm 1.93. The program progresses through time in small steps of size Δt and generates, at each time step, a random number x_{rand} chosen uniformly in the range between 0 and 1. If $r_{\text{est}}(t)\Delta t > x_{\text{rand}}$ at that time step, a spike is fired; otherwise it is not.

Algorithm 1.94. For a constant firing rate, it is faster to compute spike times t_i for $i = 1, 2, ..., n$ iteratively by generating interspike intervals from an exponential probability density(Equation 1.45). Thus we can generate spike times iteratively from the formula $t_{i+1} = t_i - \ln(x_{\text{rand}}/r)$.

Proposition 1.95. If a random variable X is uniformly distributed over the range between 0 and 1, the negative of its logarithm is exponentially distributed.

Proof. Suppose that X is uniformly distributed in $[0, 1]$, we have

$$P(X \leq x) = x.$$

Thus,

$$\begin{aligned} P(-\ln X/r \leq y) &= P(X \geq e^{-ry}) \\ &= 1 - P(X \leq e^{-ry}) = 1 - e^{-ry}, \end{aligned}$$

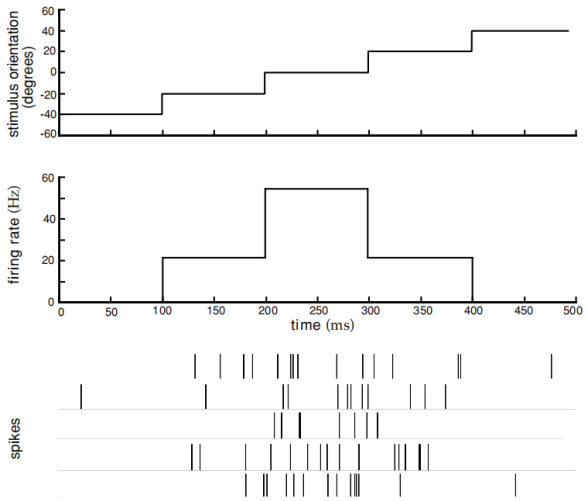
which indicates the random variable $-\ln X/r$ is follows an exponential distribution. □

Remark 1.48. The Algorithm 1.94 works only for constant firing rates. However, it can be extended to time-dependent rates by using a procedure called rejection sampling or spike thinning.

Algorithm 1.96 (Spike thinning)**.** The thinning technique requires a bound $r_{\max}$ on the estimated firing rate such that $r_{\text{est}}(t) \leq r_{\max}$ at all times. We first generate a spike sequence corresponding to the constant rate r_{max} by iterating the rule $t_{i+1} = t_i - \ln(x_{\text{rand}})/r_{\max}$. The spike are then thinned by generating another x_{rand} for each i and removing the spike at time t_i from the train if $r_{\text{est}(t_i)}/r_{\max} < x_{\text{rand}}$. If $r_{\text{est}}(t_i)/r_{\max} \geq x_{\text{rand}}$, spike i is retained. Thinning corrects for the difference between the estimated timedependent rate and the maximum rate.

Example 1.97. The following figures shows an example of a model of an orientation-selective V1 neuron constructed by Spike thinning. In this model, the estimated firing rate is determined from the response tuning curve

$$r_{est}(t) = f(s(t)) = r_{max} \exp\left(-\frac{1}{2}\left(\frac{s(t) - s_{max}}{\sigma_f}\right)^2\right). \tag{1.60}$$



This figure is the model of an orientation-selective neuron. The orientation angle (top panel) was increased from an initial value of $-40°$ by $20°$ every 100 ms. The firing rate (middle panel) was used to generate spikes (bottom panel) using a Poisson spike generator. The bottom panel shows spike sequences generated on five different trials.

1.4.5 Comparison with Data

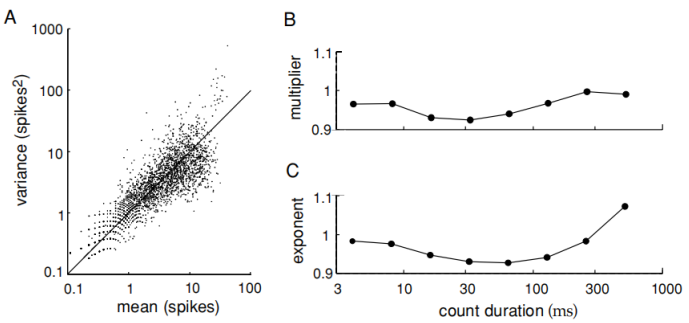
Remark 1.49. The Poisson process is simple and useful, but does it match data on neural response variability? To address this question, we examine Fano factors, interspike interval distributions, and coefficients of variation.

Remark 1.50. The Fano factor describes the relationship between the mean spike count over a given interval and the spike-count variance.

Rule 1.98 (Examine the Fano factor)**.** Mean spike counts $\langle n \rangle$ and variances σ_n^2 from a wide variety of neuronal recordings have been fitted to the Equation $\sigma_n^2 = A\langle n \rangle^B$, and the *multiplier* A and *exponent* B have been determined. The values of both A and B typically lie between 1.0 and 1.5.

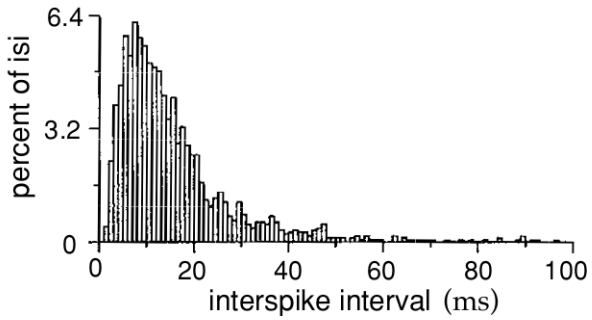
Remark 1.51. Because the Poisson model predicts $A = B = 1$, this indicates that the data show a higher degree of variability than the Poisson model would predict. However, many of these experiments involve anesthetized animals, and it is known that response variability is higher in anesthetized than in alert animals.

Example 1.99 (Fano Factor Comparision)**.** The following figures shows data for spike-count means and variances extracted from recordings of MT neurons in alert macaque monkeys using a number of different stimuli. The MT (medial temporal) area is a visual region of the primate cortex where many neurons are sensitive to image motion. The individual means and variances are scattered in figure A, but they cluster around the diagonal which is the Poisson prediction. Similarly, the results show A and B values close to 1, the Poisson values (figure B). Of course, many neural responses cannot be described by Poisson statistics, but it is reassuring to see a case where the Poisson model seems a reasonable approximation. As mentioned previously, when spike trains are not described very accurately by a Poisson model, refractory effects are often the primary reason.



Rule 1.100 (Examine interspike interval distributions)**.** Interspike interval distributions are extracted from data as interspike histograms by counting the number of intervals falling in discrete time bins.

Example 1.101 (Data Interspike Interval Distributions)**.** The following figure presents an example from the responses of a nonbursting cell in area MT of a monkey in response to images consisting of randomly moving dots with a variable amount of coherence imposed on their motion (see chapter 3 for a more detailed description).



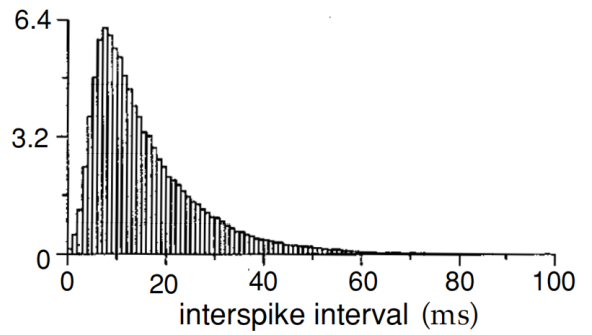
For interspike intervals longer than about 10 ms, the shape of this histogram is exponential, in agreement with Equation 1.45. However, for shorter intervals there is a discrepancy. While the homogeneous Poisson distribution of Equation 1.45 rises for short interspike intervals, the experimental results show a rapid decrease. This is the result of refractoriness making short interspike intervals less likely than the Poisson model would predict.

Proposition 1.102. The data of the Poisson model of interspike interval with a stochastic refractory period can be fitted more accurately by a gamma distribution,

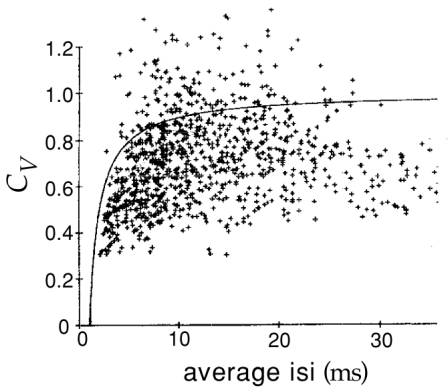
$$p[\tau] = \frac{r(r\tau)^k \exp(-r\tau)}{k!} \tag{1.61}$$

with $k > 0$, than by the exponential distribution of the Poisson model, which has $k = 0$.

Example 1.103 (Interspike Interval Distributions Model)**.** The following figure shows a theoretical histogram obtained by adding a refractory period of variable duration to the Poisson model. Spiking was prohibited during the refractory period, and then was described once again by a homogeneous Poisson process. The refractory period was randomly chosen from a Gaussian distribution with a mean of 5 ms and a standard deviation of 2 ms (only random draws that generated positive refractory periods were included). The resulting interspike interval distribution of figure 1.4.5 agrees quite well with the data.

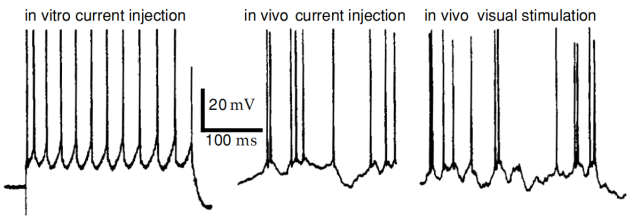


Example 1.104 (Coefficients of Variation comparision)**.** C_V values extracted from the spike trains of neurons recorded in monkeys from area MT and primary visual cortex(V1) are shown in this figure. The data have been divided into groups based on the mean interspike interval, and the coefficient of variation is plotted as a function of the mean interval, equivalent to $1/\langle r \rangle$. Except for short mean interspike intervals, the values are near 1, although they tend to cluster slightly lower than 1, the Poisson value. The small C_V values for short interspike intervals are due to the refractory period. The solid curve is the prediction of a Poisson model with refractoriness.



Remark 1.52. However, there are cases in which the accuracy in the timing and numbers of spikes fired by a neuron is considerably higher than would be implied by Poisson statistics. Furthermore, even when it successfully describes data, the Poisson model does not provide a mechanistic explanation of neuronal response variability.

Example 1.105. The following figure compares the response of V1 cells to constant current injection in vivo and in vitro. The in vitro response is a regular and reproducible spike train(left panel). The same current injection paradigm applied in vivo produces a highly irregular pattern of firing(center panel) similar to the response to a moving bar stimulus(right panel).



Remark 1.53. Although some of the basic statistical properties of firing variability may be captured by the Poisson model of spike generation, the spike generating mechanism itself in real neurons is clearly not responsible for the variability. We explore ideas about possible sources of spike-train variability in chapter 5.

Remark 1.54. Some neurons fire action potentials in clusters or bursts of spikes that can not be described by a Poisson process with a fixed rate. Bursting can be included in a Poisson model by allowing the firing rate to fluctuate in order to describe the high rate of firing during a burst. Sometimes the distribution of bursts themselves can be described by a Poisson process (such a doubly stochastic process is called a Cox process).

1.5 The Neural code

Example 1.106. Assuming that the neural response and its relation to the stimulus are completely characterized by the probability distribution of spike times as a function of

the stimulus. If spike generation can be described as an inhomogeneous Poisson process, this probability distribution can be computed from the time-dependent firing rate $r(t)$, using equation 1.37. In this case, $r(t)$ contains all the information about the stimulus that can be extracted from the spike train, and the neural code could reasonably be called a rate code.

Remark 1.55. The central issue in neural coding is whether individual action potentials and individual neurons encode independently of each other, or whether correlations between different spikes and different neurons carry significant amounts of information.

1.5.1 Independent-Spike, Independent-Neuron, and Correlation Codes

Remark 1.56. All information in this section refers to stimulating information.

Definition 1.107. A code based solely on the time-dependent firing rate is called the *independent-spike code*. This refers to the fact that the generation of each spike is independent of all the other spikes in the train.

Definition 1.108. Individual spikes do not encode independently of each other, correlations between spike times may carry additional correlation code information, which is called the *correlation codes*.

Remark 1.57. It has been found that some information is carried by correlations between two or more spikes, but this information is rarely larger than 10% of the information carried by spikes considered independently.Information could be carried by more complex relationships between spikes.Independent-spike codes are much simpler to analyze than correlation codes, and most work on neural coding assumes spike independence.

Rule 1.109. Information is typically encoded by neuronal populations.

Remark 1.58. We still consider whether individual neurons act independently, or whether correlations between different neurons carry additional information.

Remark 1.59. The analysis of population coding is easiest if the response of each neuron is considered statistically independent.It means that they can be combined without taking correlations into account.

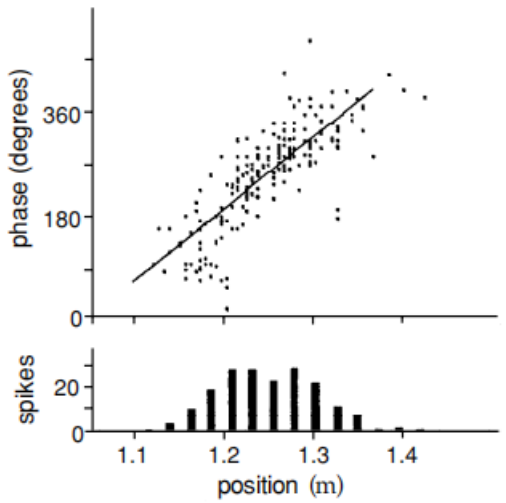
Definition 1.110. The *independent-neuron code* means the response of each neuron in a neural population is considered statistically independent.

Remark 1.60. The assumption of independent-neuron coding is a useful simplification that is not in gross contradiction with experimental data, but it is less well established and more likely to be challenged in the future than the independent-spike hypothesis.

Remark 1.61. To test the validity of independent-neuron, we should know whether correlations between the spiking of different neurons provide additional information about a stimulus that cannot be obtained by considering all of their firing patterns individually.

Principle 1.111. Synchronous firing of two or more neurons and rhythmic oscillations of population activity are mechanism for conveying information in a population correlation code.

Example 1.112. Place-cell coding of spatial location in the rat hippocampus, which at least some additional information appears to be carried by correlations between the firing patterns of neurons in a population. The firing rates of many hippocampal neurons, recorded when a rat is moving around a familiar environment, depend on the location of the animal and are restricted to spatially localized areas called the place fields of the cells. When a rat explores an environment, hippocampal neurons fire collectively in a rhythmic pattern with a frequency in the theta range, 7-12 Hz. The spiking time of an individual place cell relative to the phase of the population theta rhythm gives additional information about the location of the rat not provided by place cells considered individually.



Each dot in the upper figure shows the phase of the theta rhythm plotted against the position of the animal at the time when a spike was fired. The linear relation shows that information about position is contained in the relative phase of firing. The lower plot is a conventional place field tuning curve of spike count versus position.

1.5.2 Temporal Codes

Rule 1.113. Precise spike timing is a significant element in neural encoding. When precise spike timing or high-frequency firing-rate fluctuations are found to carry information, the neural code is often identified as a temporal code.

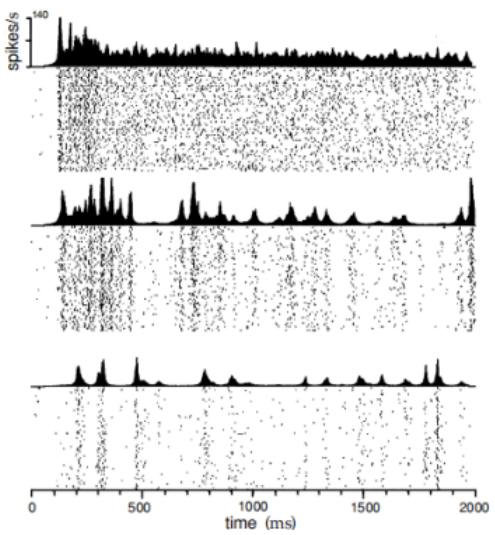
Remark 1.62. The temporal structure of a spike train or firing rate evoked by a stimulus is determined both by the dynamics of the stimulus and by the nature of the neural encoding process. Stimuli that change rapidly tend to generate precisely timed spikes and rapidly changing firing rates.

Remark 1.63. Temporal coding refers to temporal precision in the response that not only arise from the dynamics of the stimulus but also relates to properties of the stimulus.

Rule 1.114. If the independent-spike hypothesis is valid, the temporal character of the neural code is determined by the behavior of $r(t)$.

Definition 1.115. If $r(t)$ varies slowly with time, the code is typically called a *rate code*, and if it varies rapidly, the code is called *temporal code*.

Example 1.116. Different firing-rate behaviors for a neuron in area MT of a monkey recorded over multiple trials with three different stimuli(consisting of moving random dots). The activity in the top panel would typically be regarded as reflecting rate coding, and the activity in the bottom panel as reflecting temporal coding.



Remark 1.64. It is not obvious what criterion should be used to characterize the changes in $r(t)$as slow or rapid. The identification of rate and temporal coding in this way is ambiguous.

Example 1.117. Using the spikes to distinguish slow from rapid, so that a temporal code is identified when peaks in the firing rate occur with roughly the same frequency as the spikes themselves. In this case, each peak corresponds to the firing of only one, or at most a few action potentials.

Remark 1.65. When many neurons are involved, any single neuron may fire only a few spikes before its firing rate changes, but the population may produce a large number of spikes over the same time period. Thus, it is not targeted at populations.

Example 1.118. Using the stimulus to establish what makes a temporal code. In this case, a temporal code is defined as one in which information is carried by details of spike timing on a scale shorter than the fastest time characterizing variations of the stimulus. This requires that frequencies higher than those present in the stimulus.

Rule 1.119. A temporal code has been reported when using spikes to define the nature of the code, and it would be called rate codes if the stimulus were used instead.

1.6 Questions

This section states the questions that we can't solve or the concepts that we can't understand. For the first chapter, we have the following questions:

1. We already understand the article's descriptive explanation of Equation 1.6, but don't know how to derive it mathematically.

2. For Equation 1.24, we don't know how to derive the unit of the standard deviation σ_s from Equation 1.25. (Here, it said that σ_s has the units of the stimulus times the square root of the unit of time. This fact can derive from the derivation in Appendix A.)

3. We do not fully understand the derivation of Equation 1.50 from 1.49.

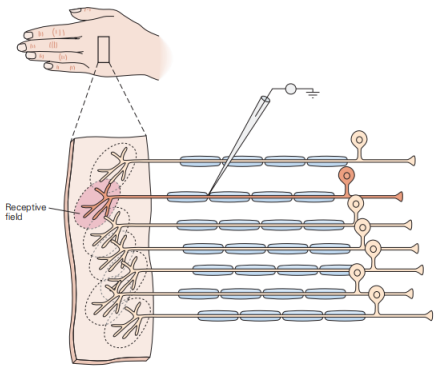
4. We have doubts about the author's interpretation of Equation 1.52, which says that if the spike times are not ordered, a change can be made to the formula. But Generally, time originally has its own sequence.

Chapter 2

Neural Encoding II: Reverse Correlation and Visual Receptive Fields

Definition 2.1. The skin area, location in the body, retinal area, or tonal domain in which stimuli can activate a sensory neuron is called its *receptive field*.

Example 2.2. The following figure is the receptive field of a touch-sensitive neuron, which denotes the region of skin where gentle tactile stimuli evoke action potentials in that neuron. Sometimes receptive fields would change over time.



Definition 2.3. *Reverse-correlation* is a technique for studying how sensory neurons add up signals from different locations in their receptive fields, and also how they sum up stimuli that they receive at different times, to generate a response.

Remark 2.1. The goal of the reverse-correlation technique is to find a function $r = f(s)$ that maps from the stimulus s to the neuronal response r, where the stimulus is a function dependent on spatial location and time $s = s(x, y, z, t)$.

Remark 2.2. The reason that this technique is called "reverse" is that we align the time origin with the neuron's response and then reverse the timeline to find what stimulus ($t < 0$) triggered the neuron's response at the current moment ($t = 0$).

Assumption 2.4. As discussed in chapter 1, sensory systems tend to adapt to the absolute intensity of a stimulus. We therefore assume throughout this chapter that the stimulus parameter $s(t)$ has been defined with its mean value subtracted out, that is,

$$\frac{1}{T}\int_0^T s(t)dt = 0. \tag{2.1}$$

2.1 Estimating Firing Rates

Remark 2.3. The response tuning curve discussed in Chapter 1 is a simple model in which firing rates were estimated as instantaneous functions of the stimulus. Nevertheless, the activity of a neuron at time t typically depends on the behavior of the stimulus over a period of time starting a few hundred milliseconds prior to t and ending perhaps tens of milliseconds before t. Reverse-correlation methods can be used to construct a more accurate model that includes the effects of the stimulus over such an extended period of time.

Remark 2.4. The **basic problem** is to construct an estimate $r_{\text{est}}(t)$ of the firing rate $r(t)$ evoked by a stimulus $s(t)$.

2.1.1 The Linear Rate Estimate

Definition 2.5. The *linear rate estimate* at any given time t is the weighted sum of the values taken by the stimulus at earlier times. With the continuous change in time, this sum actually takes the form of an integral, that is,

$$r_{\text{est}}(t) = r_0 + \int_0^\infty D(\tau)s(t-\tau)d\tau, \tag{2.2}$$

where r_0 accounts for any background firing that may occur when $s = 0$, $D(\tau)$ is a weighting factor that determines how strongly, and with what sign, the value of the stimulus at time $t - \tau$ affects the firing rate at time t.

Remark 2.5. The integral in Equation 2.2 is a linear filter.

Definition 2.6. The *error* of an estimate $r_{\text{est}}(t)$ to an actual neural response $r(t)$ is defined as

$$E = \frac{1}{T}\int_0^T (r_{\text{est}}(t) - r(t))^2 dt, \tag{2.3}$$

where T is the duration of trials.

Definition 2.7. The kernel D that minimizes the linear rate estimate error E defined in Equation 2.3 is called *optimal linear kernel* or simply called *optimal kernel*.

Proposition 2.8. The optimal kernel D satisfies

$$\int_0^\infty Q_{ss}(\tau - \tau')D(\tau')d\tau' = Q_{rs}(-\tau), \tag{2.4}$$

where $Q_{ss}(\tau) = \int_0^T s(t)s(t+\tau)/Tdt$ is the stimulus autocorrelation function, and $Q_{rs}(\tau) = \int_0^T r(t)s(t+\tau)/Tdt$ is the firing rate-stimulus correlation function, both of which were defined in chapter 1.

Proof. Using Equation 2.2 for the estimated firing rate, the expression in Equation 2.3 to be minimized is

$$E = \frac{1}{T}\int_0^T \left(r_0 + \int_0^\infty D(\tau)s(t-\tau)d\tau - r(t)\right)^2. \tag{2.5}$$

The minimum is obtained by setting the derivative of E with respect to functional derivative the function D to 0. E that depends on a function D is a functional. Finding the extrema of functionals is the subject of a branch of mathematics called the calculus of variations. A simple way to define a functional derivative is to introduce a small time interval Δt and evaluate all functions at integer multiples of Δt. We define $r_i = r(i\Delta t)$, $D_k = D(k\Delta t)$ and $s_{i-k} = s((i-k)\Delta t)$. If Δt is small enough, the integrals in Equation 2.5 can be approximated by sums,

$$E = \frac{\Delta t}{T}\sum_{i=0}^{T/\Delta t}\left(r_0 + \Delta t\sum_{k=0}^{\infty} D_k s_{i-k} - r_i\right)^2. \tag{2.6}$$

E is minimized by setting its derivative with respect to D_j for all values of j to 0,

$$\frac{\partial E}{\partial D_j} = 0 = \frac{2\Delta t}{T}\sum_{i=0}^{T/\Delta t}\left(r_0 + \Delta t\sum_{k=0}^{\infty} D_k s_{i-k} - r_i\right)s_{i-j}\Delta t. \tag{2.7}$$

Rearranging and simplifying this expression gives the condition,

$$\Delta t\sum_{k=0}^{\infty} D_k\left(\frac{\Delta t}{T}\sum_{i=0}^{T/\Delta t} s_{i-k}s_{i-j}\right) = \frac{\Delta t}{T}\sum_{i=0}^{T/\Delta t}(r_i - r_0)s_{i-j}. \tag{2.8}$$

If we take the limit $\Delta t \to 0$ and make the replacements $i\Delta t \to t$, $j\Delta t \to \tau$, and $k\Delta t \to \tau'$, the sums in Equation 2.8 turn back into integrals, the indexed variables become functions, and we find

$$\begin{aligned}&\int_0^\infty D(\tau')\left(\frac{1}{T}\int_0^T s(t-\tau')s(t-\tau)dt\right)d\tau' \\ &= \frac{1}{T}\int_0^T (r(t) - r_0)s(t-\tau)dt.\end{aligned} \tag{2.9}$$

And,

$$\begin{aligned}&\frac{1}{T}\int_0^T s(t-\tau')s(t-\tau)dt \\ &= \frac{1}{T}\int_0^T s(t-\tau+\tau-\tau')s(t-\tau)d(t-\tau) \\ &= \frac{1}{T}\int_{-\tau}^{T-\tau} s(t+\tau-\tau')s(t)dt \\ &= \frac{1}{T}\int_0^T s(t+\tau-\tau')s(t)dt = Q_{ss}(\tau-\tau'),\end{aligned}$$

where the third step follows from the translation invariance of $s(t)$. Also,

$$\begin{aligned}&\frac{1}{T}\int_0^T (r(t) - r_0)s(t-\tau)dt \\ &= \frac{1}{T}\int_0^T r(t)s(t-\tau)dt - r_0\frac{1}{T}\int_0^T s(t-\tau)dt \\ &= \frac{1}{T}\int_0^T r(t)s(t-\tau)dt = Q_{rs}(-\tau),\end{aligned}$$

where the second step follows from Assumption 2.4. Thus, Equation 2.9 can be re-expressed in the form of Equation 2.4. □

Remark 2.6. The method we are describing is a kind of reverse-correlation technique because the firing rate-stimulus correlation function is evaluated at $-\tau$ in Equation 2.4.

Definition 2.9. The *white-noise kernel* is the optimal kernel with a white-noise stimulus that satisfies $Q_{ss}(\tau) = \sigma_s^2\delta(\tau)$ with some constant σ_s.

Proposition 2.10. The *white-noise kernel* satisfies

$$D(\tau) = \frac{\langle r\rangle C(\tau)}{\sigma_s^2}, \tag{2.10}$$

where $C(\tau)$ is the spike-triggered average stimulus and $\langle r\rangle$ is the average firing rate of the neuron.

Proof. By properties of the white-noise stimulus, the left side of Equation 2.4 equals to

$$\sigma_s^2\int_0^\infty \delta(\tau-\tau')D(\tau')d\tau' = \sigma_s^2 D(\tau). \tag{2.11}$$

Thus, we have

$$D(\tau) = \frac{Q_{rs}(-\tau)}{\sigma_s^2} = \frac{\langle r\rangle C(\tau)}{\sigma_s^2}, \tag{2.12}$$

where the second step follows from the relation $Q_{rs}(-\tau) = \langle r\rangle C(\tau)$ from chapter 1. □

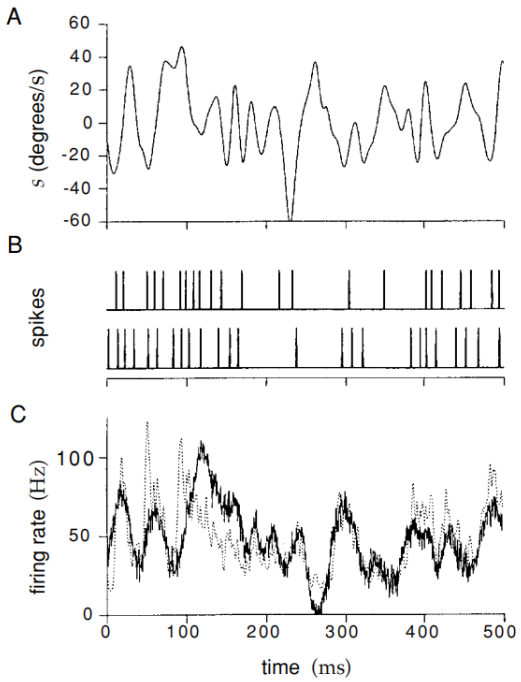
Proposition 2.11. The general solution of Equation 2.4 for an arbitrary stimulus is

$$D(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\tilde{Q}_{rs}(-\omega)}{\tilde{Q}_{ss}(\omega)}\exp\left(-i\omega\tau\right)d\omega, \tag{2.13}$$

where $\tilde{Q}_{rs}(\omega)$ and $\tilde{Q}_{ss}(\omega)$ are the Fourier transforms of $Q_{rs}(\omega)$ and $Q_{ss}(\omega)$, respectively.

Proof. The result could be obtained by the method of Fourier transforms (see *Theoretical Neuroscience* Chapter 2 Appendix A for the detailed proof). □

Example 2.12. The H1 neuron of the fly visual system responds to moving images. The following figure shows a prediction of the firing rate of this neuron obtained from a linear filter. The velocity of the moving image is plotted in A, and two typical responses are shown in B. The linear rate estimate with optimal kernel (the solid line) and the firing rate computed from the data by binning and counting spikes (the dashed line) are compared in figure C.



Remark 2.7. As the Example 2.12 shows, the linear rate estimate is a good agreement in regions where the measured rate varies slowly, but the estimate fails to capture high-frequency fluctuations of the firing rate, presumably because of nonlinear effects not captured by the linear kernel. Some such effects can be described by a static nonlinear function or including higher-order terms in a Volterra or Wiener expansion, as discussed below.

Definition 2.13. Neuronal selectivity is often characterized by describing stimuli that evoke maximal responses, subject to a constraint. This stimulus is called the *most effective stimulus.*

Remark 2.8. A constraint is essential because the linear estimate in Equation 2.2 is unbounded.

Definition 2.14. The *fixed energy constraint* is

$$\int_0^T \left(s(t')\right)^2 dt' = \text{constant}, \tag{2.14}$$

where the integral $\int_0^T \left(s(t')\right)^2 dt'$ is called *stimulus energy.*

Proposition 2.15. With the optimal kernel $D(\tau)$ and the fixed energy constraint 2.14, the most effective stimulus $s(t)$ is proportional to the optimal kernel $D(\tau)$ with

$$D(\tau) = -2\lambda s(t-\tau), \tag{2.15}$$

where $\lambda < 0$.

Proof. We impose this constraint by the method of Lagrange multipliers, which means that we must find the unconstrained maximum value with respect to s of

$$\begin{aligned} r_{\text{est}}(t) + \lambda \int_0^T s^2(t')dt' = r_0 + \int_0^\infty D(\tau)s(t-\tau)d\tau \\ + \lambda \int_0^T \left(s(t')\right)^2 dt', \end{aligned} \tag{2.16}$$

where λ is the Lagrange multiplier. Setting the derivative of this expression with respect to the function s to 0 (similar with the derivative of E in the solution to the proposition 2.8) gives Equation 2.15. □

Remark 2.9. The value of λ (which is less than 0) in Equation 2.15 is determined by requiring that condition Equation 2.14 is satisfied, but the precise value is not important for our purposes. The essential result is the proportionality between the optimal stimulus and $D(\tau)$.

Remark 2.10. The most effective stimulus analysis provides an intuitive interpretation of the linear rate estimate 2.2. At fixed stimulus energy, the integral in 2.2 measures the overlap between the actual stimulus and the most effective stimulus. In other words, it indicates how well the actual stimulus matches the most effective stimulus. Mismatches between these two reduce the value of the integral and result in lower predictions for the firing rate.

2.1.2 Volterra and Wiener Expansion

Definition 2.16. The *Volterra expansion* is the functional equivalent of the Taylor series expansion used to generate power series approximations of functions. For the case we are considering, it takes the form

$$\begin{aligned} r_{\text{est}}(t) =& r_0 + \int D(\tau)s(t-\tau)d\tau \\ &+ \iint D_2(\tau_1,\tau_2)s(t-\tau_1)s(t-\tau_2)d\tau_1 d\tau_2 \\ &+ \iiint D_3(\tau_1,\tau_2,\tau_3)s(t-\tau_1)s(t-\tau_2)s(t-\tau 3)d\tau_1 d\tau_2 d\tau_3 \\ &+ \dots. \end{aligned} \tag{2.17}$$

Definition 2.17. The series rearranged by Wiener from Equation 2.17 to make the terms easier to compute has the same first two terms of the Volterra expansion, and it is called *Wiener expansion.* The linear kernel D is called the *first Wiener kernel.*

2.1.3 Static Nonlinearities

Remark 2.11. The linear prediction has two obvious problems:

(i) There is nothing to prevent the predicted firing rate from becoming negative.

(ii) The predicted rate does not saturate, but instead increases without bound as the magnitude of the stimulus increases.

One way to deal with these and some of the other deficiencies of a linear prediction is to write the firing rate as a background rate plus a nonlinear function of the linearly filtered stimulus.

Definition 2.18. The *estimate with static nonlinearity* is

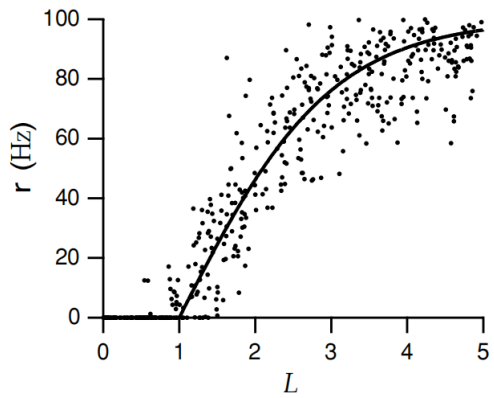
$$r_{\text{est}}(t) = r_0 + F(L(t)), \tag{2.18}$$

where F is an arbitrary function and

$$L(t) = \int_0^\infty D(\tau)s(t-\tau)d\tau. \tag{2.19}$$

F is called a *static nonlinearity* to stress that it is a function of the linear filter value evaluated instantaneously at the time of the rate estimation.

Example 2.19. F can be extracted from data by means of the graphical procedure illustrated in the following figure. First, a linear estimate of the firing rate is computed using the optimal kernel defined by Equation 2.4. Second, a plot is made of the pairs of points $(L(t), r(t))$ at various times and for various stimuli, where $r(t)$ the actual rate extracted from the data. There will be a certain amount of scatter in this plot due to the inaccuracy of the estimation. F can be extracted by fitting a function to the points on the scatter plot.



Remark 2.12. The function F typically contains constants used to set the firing rate to realistic values. These give us the freedom to normalize $D(\tau)$ in some convenient way, correcting for the arbitrary normalization by adjusting the parameters within F.

Example 2.20. The *threshold function*

$$F(L) = G[L - L_0]_+, \tag{2.20}$$

is a static nonlinearity used to introduce firing thresholds into estimates of neural responses. Here L_0 is the threshold value that L must attain before firing begins.

Remark 2.13. Above the threshold, the firing rate is a linear function of L, with G acting as the constant of proportionality. Half-wave rectification is a special case of this with $L_0 = 0$. That this function does not saturate is not a problem if large stimulus values are avoided.

Example 2.21. The *sigmoidal function*

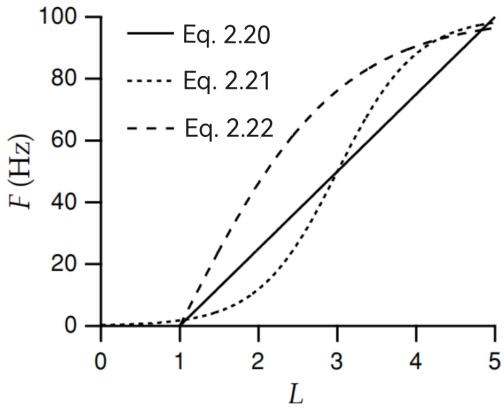
$$F(L) = \frac{r_{\max}}{1 + \exp\left(g_1(L_{1/2} - L)\right)}, \tag{2.21}$$

is a static nonlinearity used to introduce saturation into estimates of neural responses. Here $r_{\max}$ is the maximum possible firing rate, $L_{1/2}$ is the value of L for which F achieves half of this maximal value, and g_1 determines how rapidly the firing rate increases as a function of L.

Example 2.22.

$$F(L) = r_{\max}[\tanh(g_2(L - L_0))]_+ \tag{2.22}$$

is a static nonlinearity that combines a hard threshold with saturation uses a rectified hyperbolic tangent function. Here $r_{\max}$ and g_2 play similar roles as in Equation 2.21, and L_0 is the threshold.



Remark 2.14. Although the static nonlinearity can be any function, the estimate of Equation 2.18 is still restrictive because it allows for no dependence on weighted autocorrelations of the stimulus or other higher-order terms in the Volterra series.

Remark 2.15. Once the static nonlinearity is introduced, the linear kernel derived from Equation 2.4 is no longer optimal because it was chosen to minimize the squared error of the linear estimate $r_{\text{est}}(t) = r_0 + L(t)$, not the estimate with the static nonlinearity $r_{\text{est}}(t) = r_0 + F(L(t))$.

Definition 2.23. The *self-consistency condition* is that when the nonlinear estimate $r_{\text{est}}(t) = r_0 + F(L(t))$ is substituted into Equation 2.12, the relationship between the linear kernel and the firing rate-stimulus correlation function should still hold. In other words, we require that

$$D(\tau) = \frac{1}{\sigma_s^2 T}\int_0^T r_{\text{est}}(t)s(\tau - t)dt = \frac{1}{\sigma_s^2 T}\int_0^T F(L(t))s(\tau - t)dt, \tag{2.23}$$

where the second step follows from Assumption 2.4.

Theorem 2.24 (Bussgang Theorem)**.** An estimate based on the optimal kernel for linear estimation can still be self-consistent (although not necessarily optimal) when nonlinearities are present, if the stimulus used to extract the optimal kernel is Gaussian white noise.

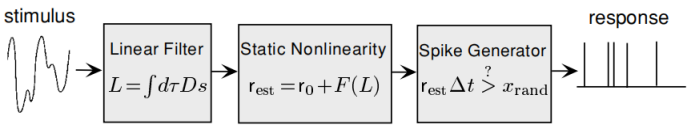
Proof. If stimulus used to extract D is Gaussian white noise, we have

$$\frac{1}{\sigma_s^2 T}\int_0^T F(L(t))s(\tau-t)dt = \frac{D(\tau)}{T}\int_0^T \frac{dF(L(t))}{dL}dt. \quad (2.24)$$

For the right side of this equation to be $D(\tau)$, the remaining expression must be equal to 1 by appropriate scaling of F. The critical identity 2.24 is based on integration by parts for a Gaussian weighted integral. □

Remark 2.16. The Bussgang Theorem suggests that Equation 2.12 will provide a reasonable kernel, even in the presence of a static nonlinearity, if the white noise stimulus used is Gaussian.

Example 2.25. A model of the spike trains evoked by a stimulus can be constructed by using the firing-rate estimate of Equation 2.18 to drive a Poisson spike generator (see chapter 1). The following figure shows the structure of such a model with a linear filter, a static nonlinearity, and a stochastic spike generator.



Remark 2.17. In some cases, the linear term fails to predict even when static nonlinearities are included and in practice including more terms in the Volterra series is quite difficult to go beyond the first few terms. We can replace the parameter s in Equation 2.19 with an appropriately chosen function of s to improve the accuracy, that is,

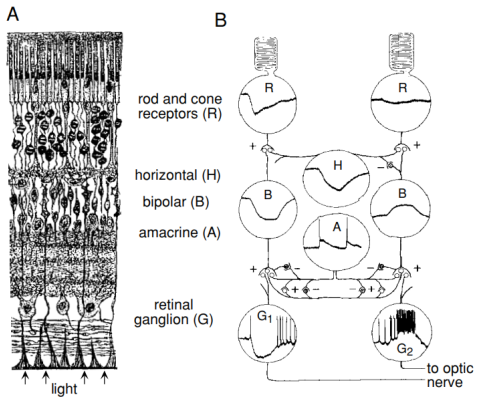
$$L(t) = \int_0^\infty D(\tau)f(s(t-\tau))d\tau.$$

A reasonable choice for this function is the response tuning curve. For time-dependent stimuli, we can think of this equation as a dynamic extension of the response tuning curve.

2.2 The Early Visual System

Principle 2.26 (Retinal Signal Conversion)**.** The conversion of a light stimulus into an electrical signal and ultimately an action potential sequence occurs in the retina. The retina is roughly composed of 3 layers of cells, *photoreceptor cells*, *bipolar cells* and *ganglion cells*. First, photoreceptor cells convert light signals into electrical signals. And then, bipolar cells are responsible for sorting and processing these electrical signals. Finally, ganglion cells will convert electrical signals into action potential sequences. In the intact eye, counterintuitively, light enters through the side opposite from the photoreceptors because vertebrate retinal cell layers are arranged in reverse order of signaling.

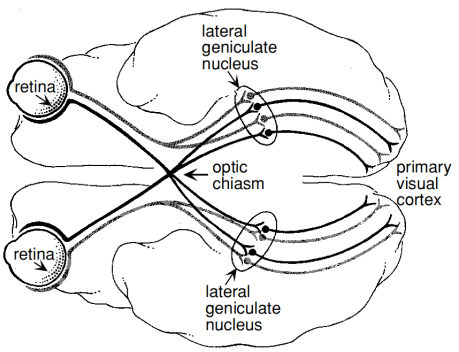
Example 2.27. The following figure A is an anatomical diagram showing the five principal cell types of the retina and figure B is a rough circuit diagram and intracellular recordings made in neurons of the retina of a mud puppy (an amphibian). The rod cells, especially the one on the left side of figure B, are hyperpolarized by the light flash. This electrical signal is passed along to bipolar and horizontal cells through synaptic connections. Note that in one of the bipolar cells, the signal has been inverted, leading to depolarization. Pluses and minuses represent excitatory and inhibitory synapses, respectively. The two retinal ganglion cells shown in the figure have different responses and transmit different sequences of action potentials. G_2 fires while the light is on, and G_1 fires when it turns off. These are called *ON* and *OFF responses*, respectively



Remark 2.18. Changing membrane potentials is adequate for signaling within the retina, where distances are small. However, it is inadequate for the task of conveying information from the retina to the brain. Thus, the ganglion cells are needed.

Definition 2.28. The output neurons of the retina are the retinal ganglion cells, whose axons form the *optic nerve*.

Principle 2.29 (Visual Pathway)**.** As the following figure shows, the optic nerve carry information from each visual hemifield up to the *optic chiasm*, where some retinal ganglion cell axons cross the midline at the optic chiasm, and then to the LGN. Cells in this nucleus send their axons along the optic radiation to the primary visual cortex.



Definition 2.30. The restricted regions of the visual field where light stimuli could active Neurons in the retina, LGN, and primary visual cortex are called *receptive fields of* the corresponding *visual neurons*.

Assumption 2.31. Patterns of illumination outside the receptive field of a given neuron cannot generate a response directly, although they can significantly affect responses to stimuli within the receptive field. We do not consider such effects, although they are of considerable experimental and theoretical interest.

Remark 2.19. Within the receptive fields, there are regions where illumination higher than the background light intensity enhances firing, and other regions where lower illumination enhances firing. The spatial arrangement of these regions determines the selectivity of the neuron to different inputs. The term *receptive field* is often generalized to refer not only to the overall region where light affects neuronal firing, but also to the spatial and temporal structure within this region.

Definition 2.32. Visually responsive neurons in the retina, LGN, and primary visual cortex are divided into two classes, depending on whether or not the contributions from different locations within the visual field sum linearly. *Simple cells* in primary visual cortex appear to satisfy this assumption. *Complex cells* in primary visual cortex do not show linear summation across the spatial receptive field, and nonlinearities must be included in descriptions of their responses.

Assumption 2.33. To streamline the discussion in this chapter, we consider only gray-scale images, although the methods presented can be extended to include color. We also restrict the discussion to two-dimensional visual images, ignoring how visual responses depend on viewing distance and encode depth.

Remark 2.20. In discussing the response properties of retinal, LGN, and V1 neurons, we do not follow the path of the visual signal, nor the historical order of experimentation, but instead begin with primary visual cortex and then move back to the LGN and retina. And the emphasis of this chapter is on properties of individual neurons.

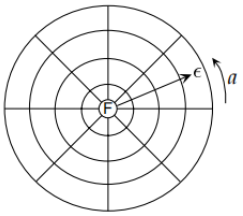
2.2.1 The Retinotopic Map

Definition 2.34. The *retinotopic map* is a map from the visual world to the cortical surface that make sure neighboring points in a visual image evoke activity in neighboring regions of visual cortex.

Remark 2.21. A striking feature of most visual areas in the brain, including primary visual cortex, is that the visual world is mapped onto the cortical surface in this topographic manner. The retinotopic map refers to the transformation from the coordinates of the visual world to the corresponding locations on the cortical surface.

Definition 2.35. The image point that focuses onto the fovea or center of the retina is called the *fixation point*.

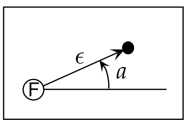
Definition 2.36. Locations on a sphere can be represented using the same longitude and latitude angles used for the surface of the earth, which called *spherical coordinate system*.



The north pole is located at the fixation point, the latitude coordinate is called the *eccentricity* ϵ, and the longitude coordinate, measured from the horizontal meridian, is called the *azimuth* a.

Principle 2.37. In primary visual cortex, the visual world is split in half, with the region $-90^\circ \leq a \leq 90^\circ$ for ϵ from 0° to about 70° (for both eyes) represented on the left side of the brain, and the reflection of this region about the vertical meridian represented on the right side of the brain.

Definition 2.38. The *polar coordinate system* used to parameterize image location is shown in the following figure.



The rectangle represents a *tangent screen*, the filled circle is the location of a particular image point on the screen, the origin of the polar coordinate system is the fixation point F, the *eccentricity* ϵ is proportional to the radial distance from the fixation point to the image point, and a is the angle between the radial line from F to the image point and the horizontal axis.

Remark 2.22. In most experiments, images are displayed on a tangent screen that does not coincide exactly with the sphere discussed in Definition 2.36. However, if the tangent screen is not too large, the difference is negligible, and the eccentricity and azimuth angles approximately coincide with polar coordinates on the screen.

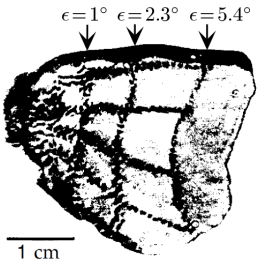
Assumption 2.39. In the following, the eccentricity ϵ and the x and y coordinates of the Cartesian system that are based on measuring distances on the screen are converted to degrees by

$$\frac{l}{r} \times \frac{180^\circ}{\pi}, \tag{2.25}$$

where l is the distance on the screen, r is the distance from the eye to the screen.

Remark 2.23. Assumption 2.39 makes sense because it is the angular, not the absolute size and location of an image that is typically relevant for studies of the visual system. And Equation 2.25 is similar to the arc length-radian relationship.

Example 2.40. The following figure shows an autoradiograph of the posterior region of the primary visual cortex from the left side of a macaque monkey brain. The pattern is a radioactive trace of the activity evoked by an image like that in the figure of Definition 2.36. The vertical lines correspond to circles at eccentricities of 1°, 2.3°, and 5.4°, and the horizontal lines (from top to bottom) represent radial lines in the visual image at a values of −90°, −45°, 0°, 45°, and 90°. Only the part of cortex corresponding to the central region of the visual field on one side is shown.



Assumption 2.41. To construct the retinotopic map, we assume that eccentricity is mapped onto the horizontal coordinate X of the cortical sheet, and a is mapped onto its Y coordinate.

Definition 2.42. The *cortical magnification factor* determines the distance across a flattened sheet of cortex separating the activity evoked by two nearby image points.

Assumption 2.43. We assume the cortical magnification factor is isotropic, denoted by $M(\epsilon)$.

Proposition 2.44. X and Y as the functions of ϵ and a satisfy

$$\frac{dX}{d\epsilon} = M(\epsilon), \tag{2.26}$$

and

$$\frac{dY}{da} = -\frac{\epsilon\pi}{180^\circ} M(\epsilon). \tag{2.27}$$

Proof. Suppose that there are two image points in question (ϵ, a) and $(\epsilon+\Delta\epsilon, a)$, the angular distance between these two points is $\Delta\epsilon$, and the distance separating the activity evoked by these two image points on the cortex is ΔX. By the definition of $M(\epsilon)$, these two quantities satisfy $\Delta X = M(\epsilon)\Delta\epsilon$ or, taking the limit as ΔX and $\Delta\epsilon$ go to 0,

$$\frac{dX}{d\epsilon} = M(\epsilon).$$

Suppose that there are the other two image points in question (ϵ, a) and $(\epsilon, a+\Delta a)$, the angular distance between these two points is

$$\Delta a \times \frac{\epsilon\pi}{180^\circ},$$

where ϵ corrects for the increase of arc length as a function of eccentricity, and $\frac{\pi}{180^\circ}$ converts from degrees to radians. The separation on the cortex ΔY corresponding to these points satisfies $\Delta Y = \Delta a \frac{\epsilon\pi}{180^\circ} M(\epsilon)$. Taking the limit $\Delta a \to 0$,

$$\frac{dY}{da} = -\frac{\epsilon\pi}{180^\circ} M(\epsilon).$$

The minus sign in this relationship appears because the visual field is inverted on the cortex. □

Example 2.45. The cortical magnification factor for the macaque monkey, obtained from results such as the figure in Example 2.40, is approximately

$$M(\epsilon) = \frac{\lambda}{\epsilon_0 + \epsilon}, \tag{2.28}$$

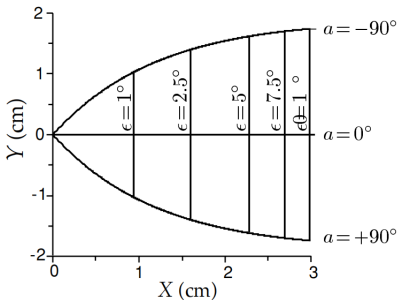
with $\lambda \approx 12$ mm and $\epsilon_0 \approx 1^\circ$. Integrating Equation 2.26 and defining $X = 0$ to be the point representing $\epsilon = 0$, we find

$$X = \lambda \ln(1 + \frac{\epsilon}{\epsilon_0}). \tag{2.29}$$

Similarly,

$$Y = -\frac{\lambda\epsilon a\pi}{(\epsilon_0 + \epsilon)180^\circ}. \tag{2.30}$$

The following figure shows that these coordinates agree fairly well with the map in Example 2.40.



Example 2.46. For $\epsilon \gg 1^\circ$, equations 2.29 and 2.30 reduce to

$$X \approx \lambda \ln(\frac{\epsilon}{\epsilon_0}), Y \approx -\frac{\lambda\pi a}{180^\circ}.$$

These two formulas can be combined by defining the complex numbers $Z = X + iY$ and $z = \frac{\epsilon}{\epsilon_0} \exp(-i\pi a/180^\circ)$, and writing

$$Z = \lambda \ln(z).$$

For this reason, the cortical map is sometimes called a *complex logarithmic map*. For an image scaled radially by a factor γ, eccentricities change according to $\epsilon \to \gamma\epsilon$ while a is unaffected. Scaling of the eccentricity produces a shift

$$X \to X + \lambda \ln(\gamma)$$

over the range of values where the simple logarithmic form of the map is valid. The logarithmic transformation thus causes images that are scaled radially outward on the retina to be represented at locations on the cortex translated in the X direction.

Remark 2.24. For smaller ϵ, the map we have derived is only approximate even in the complete form given by equations 2.29 and 2.30. This is because the cortical magnification factor is not really isotropic, as we have assumed in this derivation, and a complete description requires accounting for the curvature of the cortical surface.

2.2.2 Visual Stimuli

Remark 2.25. Pixel locations are parameterized by Cartesian coordinates x and y. However, pixel-by-pixel light intensities are not a useful way of parameterizing a visual image for the purposes of characterizing neuronal responses. This is because visually responsive neurons, like many sensory neurons, adapt to the overall level of screen illumination.

Definition 2.47. We describe the *visual stimulus* by a function $s(x, y, t)$ that is proportional to the difference between the luminance at the point (x, y) at time t and the average or background level of luminance.

Remark 2.26. Definition 2.47 could avoid dealing with adaptation effects.

Definition 2.48. The *contrast* is the resulting quantity that $s(x, y, t)$ divided by the background luminance level, making it dimensionless.

Definition 2.49. A commonly used stimulus, the *counterphase sinusoidal grating*, is described by

$$s(x, y, t) = A\cos(Kx\cos\Theta + Ky\sin\Theta - \Phi)\cos(\omega t), \quad (2.31)$$

where K and ω are the *spatial* and *temporal frequencies* of the grating (these are angular frequencies), Θ is its *orientation*, Φ is its *spatial phase*, and A is its *contrast amplitude*.

Example 2.50. The following figure shows a similar grating (a spatial square wave is drawn rather than a sinusoid) and illustrates the significance of the parameters in Definition 2.49. The lighter stripes are regions where $s > 0$, and $s < 0$ within the darker stripes. This stimulus oscillates in both space and time:

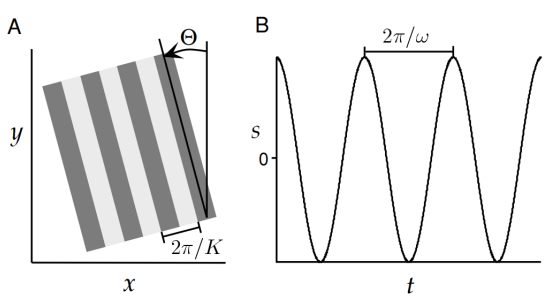
(i) At any fixed time, it oscillates in the direction perpendicular to the orientation angle Θ as a function of position, with wavelength $\frac{2\pi}{K}$ (figure A). A stimulus with $\Theta = 0$ varies in the x direction.

(ii) At any fixed position, it oscillates in time with period $\frac{2\pi}{\omega}$ (figure B).

Changing Φ by an amount $\Delta\Phi$ shifts the grating in the direction perpendicular to its orientation by a fraction $\frac{\Delta\Phi}{2\pi}$ of its wavelength, that is, $\frac{\Delta\Phi}{K}$, derived from

$$\begin{aligned} &Kx\cos\Theta + Ky\sin\Theta - (\Phi + \Delta\Phi) \\ &= Kx\cos\Theta + Ky\sin\Theta - \Delta\Phi(\sin\Theta^2 + \cos\Theta^2) - \Phi \\ &= K(x - \frac{\Delta\Phi}{K}\cos\Theta)\cos\Theta + K(y - \frac{\Delta\Phi}{K}\sin\Theta)\sin\Theta - \Phi. \end{aligned}$$

The contrast amplitude A controls the maximum degree of difference between light and dark areas.



Exercise 2.51. Prove that units of parameters in Definition 2.49 are as follows:

parameter	unit
K	radians per degree
$\frac{K}{2\pi}$	cycles per degree
Φ	radians
ω	radians/s(second)
$\frac{\omega}{2\pi}$	HZ

Remark 2.27. Experiments that consider reverse correlation and spike-triggered averages use various types of random and white-noise stimuli in addition to bars and gratings.

Definition 2.52. A *white-noise image* is one visual stimulus that is uncorrelated in both space and time so that

$$\frac{1}{T}\int_0^T s(x, y, t)s(x', y', t + \tau)dt = \sigma_s^2\delta(\tau)\delta(x - x')\delta(y - y'). \quad (2.32)$$

Remark 2.28. In practice a discrete approximation of such a stimulus must be used by dividing the image space into pixels and time into small bins. In addition, more structured random sets of images (randomly oriented bars, for example) are sometimes used to enhance the responses obtained during stimulation.

2.2.3 The Nyquist Frequency

Remark 2.29. Many factors limit the maximal spatial frequency that can be resolved by the visual system, one interesting effect arises from the size and spacing of individual photoreceptors on the retina. The region of the retina with the highest resolution is the fovea at the center of the visual field. Within the macaque or human fovea, cone photoreceptors are densely packed in a regular array.

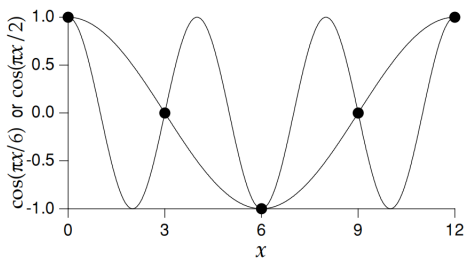
Definition 2.53. Along any direction in the visual field, a regular array of tightly packed photoreceptors of size Δx samples points at locations $m\Delta x$ for $m = 1, 2, \ldots$. The (angular) frequency that defines the *resolution* of such an array is called the *Nyquist frequency* and is given by

$$K_{\text{nyq}} = \frac{\pi}{\Delta x}. \quad (2.33)$$

Example 2.54. Consider sampling two cosine gratings with spatial frequencies of K and $2K_{\text{nyq}} - K$, with $K < K_{\text{nyq}}$. These are described by $s = \cos(Kx)$ and $s = \cos((2K_{\text{nyq}} - K)x)$. At the sampled points $m\Delta x$, these functions are identical because

$$\cos((2K_{\text{nyq}} - K)m\Delta x) = \cos(2\pi m - Km\Delta x) = \cos(Km\Delta x),$$

which follows from the periodicity and evenness of the cosine function (see the following figure).



As a result, these two gratings cannot be distinguished by examining them only at the sampled points.

Remark 2.30 (The importance of Nyquist frequency)**.** As discussed in Example 2.54, any two spatial frequencies $K < K_{\text{nyq}}$ and $2K_{\text{nyq}} - K$ can be confused with one another in this way, a phenomenon known as aliasing. Conversely, if an image is constructed solely of frequencies less than K_{nyq}, it can be reconstructed perfectly from the finite set of samples provided by the array. (Note that, images with smaller K will be easier to distinguish because of their bigger wavelengths.)

Example 2.55. There are 120 cones per degree at the fovea of the macaque retina, which makes $\Delta x = 1/120$ and

$$\frac{K_{\text{nyq}}}{2\pi} = \frac{1}{2\Delta x} = 60 \text{ cycles per degree.}$$

In this result, we have divided the right side of Equation 2.33, which gives K_{nyq} in units of radians per degree, by 2π to convert the answer to cycles per degree.

2.3 Reverse-Correlation Methods: Simple Cells

Definition 2.56. Given the light intensity of a visual image $s(x, y, t)$ and a spike sequence $\{t_i\}_{i=1}^n$, the *spike-triggered average stimulus* is a function of three variables

$$C(x, y, \tau) = \frac{1}{\langle n \rangle} \left\langle \sum_{i=1}^{n} s(x, y, t_i - \tau) \right\rangle, \tag{2.34}$$

where the brackets denote trial averaging, and we have used the approximation $1/n \approx 1/\langle n \rangle$.

Definition 2.57. The *correlation function* between the firing rate at time t and the stimulus at time $t + \tau$, for trials of duration T is defined as

$$Q_{rs}(x, y, \tau) = \frac{1}{T} \int_0^T r(t) s(x, y, t + \tau) dt. \tag{2.35}$$

Proposition 2.58. The spike-triggered average is related to the reverse-correlation function by

$$C(x, y, \tau) = \frac{Q_{rs}(x, y, -\tau)}{\langle r \rangle}, \tag{2.36}$$

where $\langle r \rangle = \langle n \rangle / T$ is as usual, the average firing rate over the entire trial.

Proof. The proof is similar with the one in Chapter 1. □

Remark 2.31. To estimate the firing rate of a neuron in response to a particular image, we add a function of the output of a linear filter of the stimulus to the background firing rate r_0, as in Equation 2.18, $r_{\text{est}}(t) = r_0 + F(L(t))$. Because visual stimuli depend on spatial location, we must decide how contributions from different image locations are to be combined to determine $L(t)$. Note that, firing rates are not a function of x and y.

Definition 2.59. Suppose that the contributions from linear response estimate different spatial points sum linearly, the *linear response estimate* $L(t)$ is obtained by integrating over all x and y values:

$$L(t) = \int_0^\infty \iint D(x, y, \tau) s(x, y, t - \tau) dx dy d\tau, \tag{2.37}$$

where the kernel $D(x, y, \tau)$ determines how strongly, and with what sign, the visual stimulus at the point (x, y) and at time $t - \tau$ affects the firing rate of the neuron at time t.

Proposition 2.60. The optimal kernel is given in terms of the firing rate-stimulus correlation function, or the spike-triggered average, for a white-noise stimulus with variance parameter σ_s^2 by

$$D(x, y, \tau) = \frac{Q_{rs}(x, y, -\tau)}{\sigma_s^2} = \frac{\langle r \rangle C(x, y, \tau)}{\sigma_s^2}. \tag{2.38}$$

Proof. The proof is similar with the one in Chapter 1. □

Definition 2.61. The kernel $D(x, y, \tau)$ defines the *space-time receptive field* of a neuron.

Remark 2.32. Because $D(x, y, \tau)$ is a function of three variables, it can be difficult to measure and visualize.

Definition 2.62. If the spatial structure of one neuron's receptive field does not change over time except by an overall multiplicative factor, the kernel can be written as a product of two functions, one that describes the *spatial receptive field* and the other, the *temporal receptive field*,

$$D(x, y, \tau) = D_s(x, y) D_t(\tau). \tag{2.39}$$

Such neurons are said to have *separable space-time receptive field.*

Definition 2.63. When $D(x, y, \tau)$ cannot be written as the product of two terms, the neuron is said to have a *nonseparable space-time receptive field.*

Assumption 2.64 (Normalization)**.** Given the freedom in Equation 2.18 to set the scale of D (by suitably adjusting the function F), we typically normalize D_s so that its integral is 1, and use a similar rule for the components from which D_t is constructed, that is,

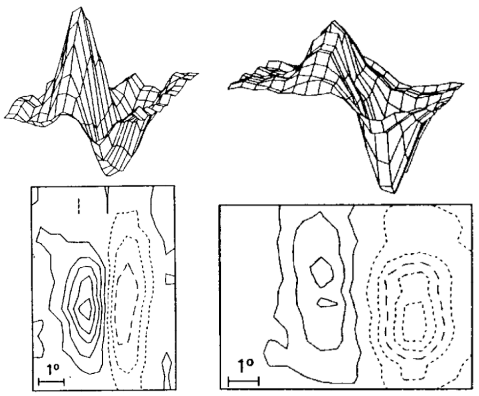
$$\iint D_s(x, y) dx dy = 1 \text{ and } \int_0^\infty \mathrm{D_t}(\tau) \mathrm{d}\tau = 1.$$

Remark 2.33. We begin our analysis by studying first the spatial and then the temporal components of a separable space-time receptive field, and then proceed to the nonseparable case. For simplicity, we ignore the possibility that cells can have slightly different receptive fields for the two eyes, which underlies the disparity tuning considered in chapter 1.

2.3.1 Spatial Receptive Fields

Definition 2.65. Regions within the receptive field where D_s is positive, is called *ON regions*, and where it is negative, is called *OFF regions*.

Example 2.66. The following figures show the spatial structure of the receptive fields of two neurons in cat primary visual cortex determined by averaging stimuli between 50 ms and 100 ms prior to an action potential, that is, $\int_{50}^{100} C(x, y, \tau)d\tau/50$.



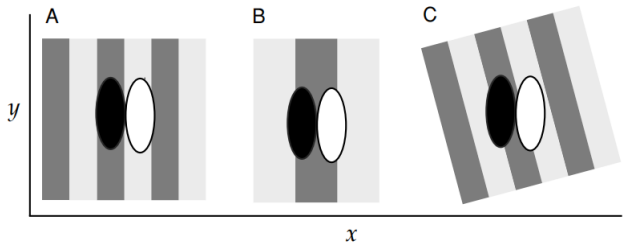
The upper plots are three-dimensional representations, with the horizontal dimensions acting as the x-y plane and the vertical dimension indicating the magnitude and sign of $D_s(x, y)$. The lower contour plots represent the x-y plane. Regions with solid contour curves are ON areas where $D_s(x, y) > 0$, and regions with dashed contours are OFF areas where $D_s(x, y) < 0$.

Proposition 2.67. The response of a neuron is enhanced if ON regions are illuminated ($s > 0$) or if OFF regions are darkened ($s < 0$) relative to the background level of illumination. Conversely, they are suppressed by darkening ON regions or illuminating OFF regions.

Proof. The integral of the linear kernel times the stimulus can be visualized by noting how the OFF and ON regions overlap the image. □

Remark 2.34. From Proposition 2.67, the neurons of figures in Example 2.66 respond most vigorously to light-dark edges positioned along the border between the ON and OFF regions, and oriented parallel to this border and to the elongated direction of the receptive fields, shown below.

Example 2.68. Grating stimuli superimposed on spatial receptive fields similar to those shown in Example 2.66. The receptive field is shown as two oval regions, one dark to represent an OFF area where $D_s < 0$ and one white to denote an ON region where $D_s > 0$.



Remark 2.35. The above examples show receptive fields with two major subregions. Simple cells are found with from one to five subregions. Along with the ON-OFF patterns we have seen, another typical arrangement is a three-lobed receptive field with OFF-ON-OFF or ON-OFF-ON subregions.

Definition 2.69. A *Gabor function* is a product of a Gaussian function and a sinusoidal function.

Definition 2.70. If the coordinates x and y are chosen such that the borders between the ON and OFF regions are parallel to the y axis and the origin of the coordinates is placed at the center of the receptive field, the Gabor function

$$D_s(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(kx - \phi) \quad (2.40)$$

can be used to approximate the observed receptive field structures. The parameters in this function determine the properties of the spatial receptive field:

(i) σ_x and σ_y, the *receptive field size*, determine its extent in the x and y directions, respectively.

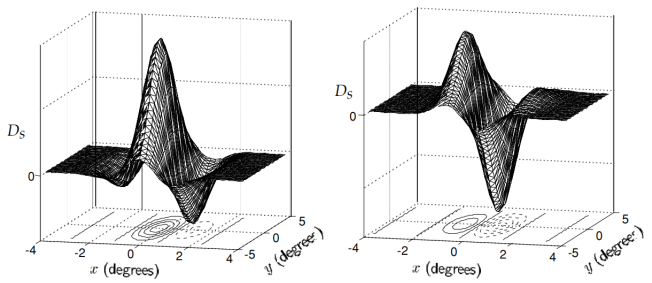
(ii) k, the *preferred spatial frequency*, determines the spacing of light and dark bars that produce the maximum response (the preferred spatial wavelength is $2\pi/k$).

(iii) ϕ, the *preferred spatial phase*, determines where the ON-OFF boundaries fall within the receptive field.

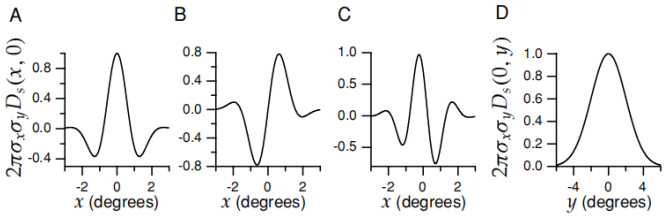
Proposition 2.71. For the spatial receptive field Equation 2.40, the sinusoidal grating described by Equation 2.31 that produces the maximum response for a fixed value of A has $K = k$, $\Phi = \phi$, and $\Theta = 0$.

Notation 13. The Gabor functions mentioned below refer to Equation 2.40.

Example 2.72. The Gabor functions chosen specifically to match the data in Example 2.66. These two figures are plotted with $\sigma_x = 1°$, $\sigma_y = 2°$, $1/k = 0.56°$ and $\phi = 1 - \pi/2$ (left), $\phi = 1 - \pi$ (right).



Example 2.73. x- and y- plots of a variety of Gabor functions (Equation 2.40) with different parameter values. For convenience, these plots are the dimensionless function $2\pi\sigma_x\sigma_y D_s$.



Responding parameters are as follows:

(i) A with $\sigma_x = 1°$, $1/k = 0.5°$, $\phi = 0$ and $y = 0$ is symmetric about $x = 0$.

(ii) B with $\sigma_x = 1°$, $1/k = 0.5°$, $\phi = \pi/2$ and $y = 0$ is antisymmetric about $x = 0$ and corresponds to using a sine instead of a cosine function in Equation 2.40.

(iii) C with $\sigma_x = 1°$, $1/k = 0.33°$, $\phi = \pi/4$ and $y = 0$ has no particular symmetry properties with respect to $x = 0$.

(iv) D with $\sigma_y = 2°$ and $x = 0$ is simply a Gaussian.

Remark 2.36. As seen in Example 2.73, Gabor functions can have various types of symmetry, and variable numbers of significant oscillations (or subregions) within the Gaussian envelope.

Remark 2.37. The response characterized by Equation 2.40 is maximal if light-dark edges are parallel to the y axis, so the preferred orientation orientation for a stimulus is 0.

Proposition 2.74. An arbitrary preferred orientation θ can be generated by rotating the coordinates, making the substitutions $x \to a\cos(\theta) + y\sin(\theta)$ and $y \to y\cos(\theta) - x\sin(\theta)$ in Equation 2.40. This produces a spatial receptive field that is maximally responsive to a grating with $\Theta = \theta$.

Proof. Actually, this rotation is from the rotation matrix

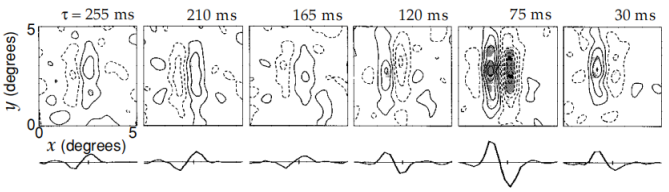
$$M(\theta) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix},$$

which rotates a vector in two-dimensional space by θ clockwise. □

Proposition 2.75. Similarly, a receptive field centered at the point (x_0, y_0) rather than at the origin can be constructed by making the substitutions $x \to x - x_0$ and $y \to y - y_0$, where the (x_0, y_0) is called the *receptive field center*.

2.3.2 Temporal Receptive Fields

Example 2.76. The following figure reveals the temporal development of the space-time receptive field of a neuron in the cat primary visual cortex through a series of snapshots of its spatial receptive field. Each panel is a plot of $D(x, y, \tau)$ for a different value of τ. The curves below the contour diagrams are one-dimensional plots of the receptive field as a function of x alone. Around $\tau = 210$ ms, a two-lobed OFF-ON receptive field is evident. As τ decreases, this structure first fades away and then reverses, so that the receptive field at $\tau = 75$ ms has the opposite sign from what appeared at $\tau = 210$ ms.



The stimulus preferred by this cell is thus an appropriately aligned dark-light boundary that reverses to a light-dark boundary over time.

Remark 2.38. In the above figure, for $\tau > 300$ ms, there is little correlation between the visual stimulus and the upcoming spike. Due to latency effects, the spatial structure of the receptive field is less significant for $\tau < 75$ ms.

Proposition 2.77. The neuron in Example 2.76 has approximately a separable space-time receptive field.

Proof. Although the magnitudes and signs of the different spatial regions vary over time, their locations and shapes remain fairly constant. This indicates that the neuron has, to a good approximation, a separable space-time receptive field. □

Definition 2.78. The phenomenon that a neuron's receptive field reverses with time is called the *reversal effect*.

Remark 2.39. Reversal effects are a common feature of space-time receptive fields.

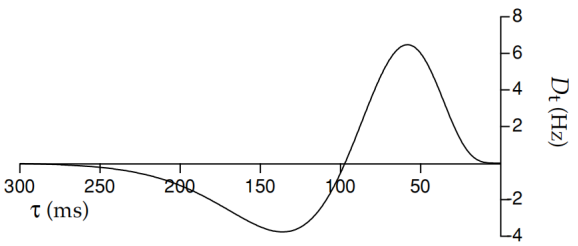
Proposition 2.79. The reversal can be described by a function

$$D_t(\tau) = \begin{cases} \alpha \exp(-\alpha\tau)\left(\frac{(\alpha\tau)^5}{5!} - \frac{(\alpha\tau)^7}{7!}\right) & \tau \geq 0, \\ 0 & \tau < 0. \end{cases} \tag{2.41}$$

Here, α is a constant that sets the scale for the temporal development of the function.

Proof. The function $D_t(\tau)$ of Equation 2.41 rises from 0, becomes positive, then negative, and ultimately returns to 0 as τ increases. □

Example 2.80. Temporal structure of a receptive field. (Plot $D_t(\tau)$ in Equation 2.41 with $\alpha = 1/(15$ ms$)$.)



Remark 2.40. Single-phase responses are also seen for V1 neurons, and these can be described by eliminating the second term in Equation 2.41. Three-phase responses, which are sometimes seen, must be described by a more complicated function.

2.3.3 Response of a Simple Cell to a Counterphase Grating

Remark 2.41. The response of a simple cell to a counterphase grating stimulus (Equation 2.31) can be estimated by computing the function $L(t)$.

Proposition 2.81. For the separable receptive field, the linear estimate of the response can be written as the product of two terms,

$$L(t) = L_s L_t(t), \tag{2.42}$$

where

$$L_s = \iint D_s(x,y) A \cos\left(Kx\cos(\Theta) + Ky\sin(\Theta) - \Phi\right) dxdy \tag{2.43}$$

and

$$L_t(t) = \int_0^\infty D_t(\tau)\cos(\omega(t-\tau))d\tau. \tag{2.44}$$

Assumption 2.82. We assume that $D_s(x,y)$ and $D_t(\tau)$ used in this section are from Equation 2.40 and 2.41, respectively.

Exercise 2.83. Compute these integrals in equations 2.43 and 2.44 for $D_s(x,y)$ in Equation 2.40 with $\sigma_x = \sigma_y = \sigma$ and $D_t(\tau)$ in Equation 2.41.

Proposition 2.84. If the spatial phase of the stimulus and the preferred spatial phase of the receptive field are 0 ($\Phi = \phi = 0$),

$$L_s = A\exp\left(-\frac{\sigma^2(k^2+K^2)}{2}\cosh(\sigma^2 kK\cos(\Theta))\right), \tag{2.45}$$

which determines the orientation and spatial frequency tuning for an optimal spatial phase.

Lemma 2.85. For a grating with the preferred orientation $\Theta = 0$ and a spatial frequency that is not too small, the full expression for L_s can be simplified by noting that $\exp(-\sigma^2 kK) \approx 0$ for the values of $k\sigma$ noemally encountered.

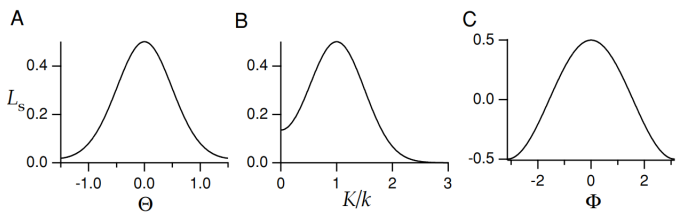
Example 2.86. If $K = k$ and $k\sigma = 2$, $\exp(-\sigma^2 kK) = 0.02$.

Proposition 2.87. Using the approximation in Lemma 2.85,

$$L_s = \frac{A}{2}\exp\left(-\frac{\sigma^2(k-K)^2}{2}\right)\cos(\phi - \Phi), \tag{2.46}$$

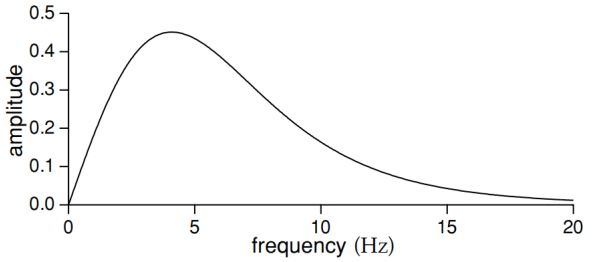
which reveals a Gaussian dependence on spatial frequency and a cosine dependence on spatial phase.

Example 2.88. Selectivity of a Gabor filter (Equation 2.43 with $D_s(x,y)$ from Equation 2.40) with $\theta = \phi = 0$, $\sigma_x = \sigma_y = \sigma$, and $k\sigma = 2$ acting on a cosine grating with $A = 1$.



(A) L_s as a function of stimulus orientation Θ for a grating with the preferred spatial frequency and phase, $K = k$ and $\Phi = 0$. (B) L_s as a function of the ratio of the stimulus spatial frequency to its preferred value, K/k, for a grating oriented in the preferred direction $\Theta = 0$ and with the preferred phase $\Phi = 0$. (C) L_s as a function of stimulus spatial phase Φ for a grating with the preferred spatial frequency and orientation, $K = k$ and $\Theta = 0$.

Example 2.89. The temporal frequency dependence of the amplitude of the linear response estimate is plotted as a function of the temporal frequency of the stimulus ($\omega/2\pi$ rather than the angular frequency ω).



The peak value around 4 Hz and roll-off above 10 Hz are typical for V1 neurons and for cortical neurons in other primary sensory areas as well.

2.3.4 Space-Time Receptive Fields

Remark 2.42. To display the function $D(x,y,\tau)$ in a space-time plot rather than as a sequence of spatial plots , we suppress the y dependence and plot an x-τ projection of the space-time kernel.

Example 2.90 (A separable space-time receptive field)**.** Figure A shows a space-time plot of the receptive field of a simple cell in the cat primary visual cortex. This receptive field is approximately separable, and it has side-by-side OFF and ON regions that reverse as a function of τ.

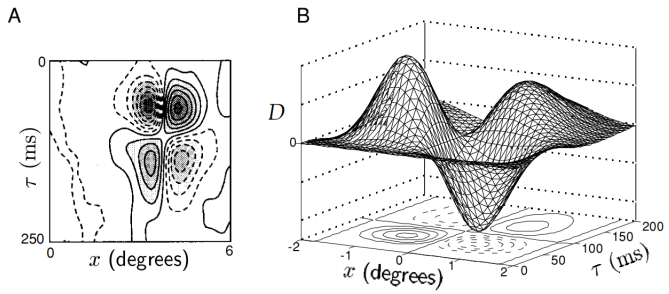
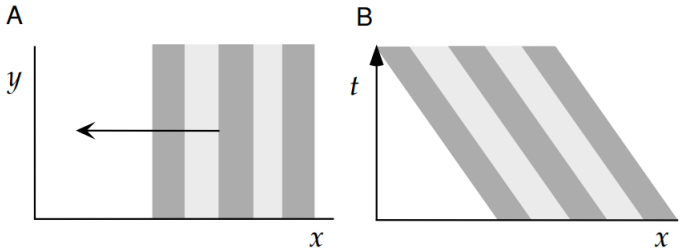


Figure B shows an x-τ plot of a separable space-time kernel, similar to the one in A, generated by multiplying a Gabor function (evaluated at $y = 0$) with $\sigma_x = 1°$, $1/k = 0.56°$ and $\phi = \pi/2$ by the temporal kernel of Equation 2.41 with $1/\alpha = 15$ ms.

Remark 2.43. We can also plot the visual stimulus in a space-time diagram, suppressing the y coordinate by assuming that the image does not vary as a function of y.

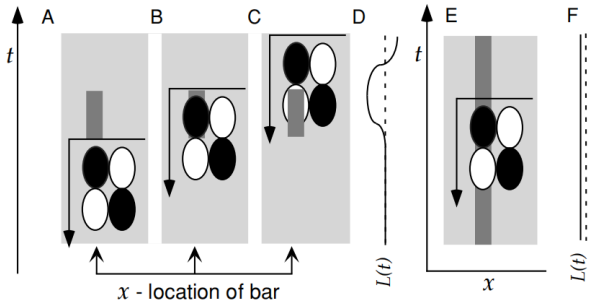
Example 2.91 (Space and space-time diagrams of a moving grating)**.** Figure A shows a grating of vertically oriented stripes moving to the left on an x-y plot. In the x-t plot of figure B, this image appears as a series of sloped dark and light bands. These represent the projection of the image in A onto the x axis evolving as a function of time. The leftward slope of the bands corresponds to the leftward movement of the image.



Principle 2.92. Most neurons in primary visual cortex do not respond strongly to static images, but respond vigorously to flashed and moving bars and gratings.

Remark 2.44. The receptive field structure of Example 2.90 reveals why this is the case in Principle 2.92.

Example 2.93 (Why a flashed bar is a effective stimulus)**.** The following figures show responses to dark bars estimated from a separable space-time receptive field.



The linear estimate of the response at any time is determined by positioning the receptive field diagram so that its horizontal axis matches the time of response estimation and noting how the OFF and ON regions overlap with the image.

(A-C) The image is a dark bar that is flashed on for a short interval of time. There is no response (figure A) until the dark image overlaps the OFF region (figure B) when $L(t) > 0$. The response is later suppressed when the dark bar overlaps the ON region (figure C) and $L(t) < 0$.

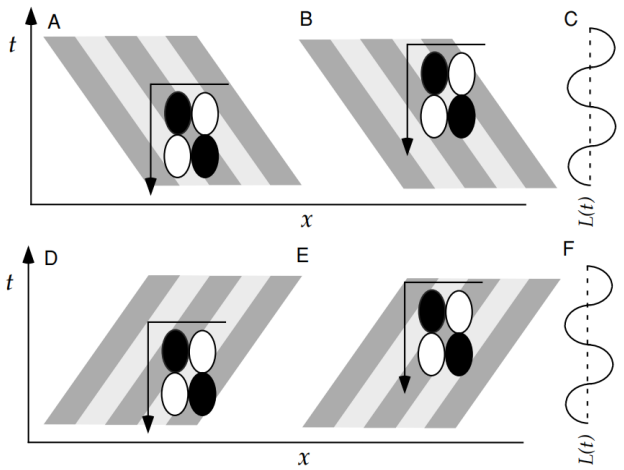
(D) A plot of $L(t)$ versus time corresponding to the responses generated in figures A-C. Time runs vertically in this plot, and $L(t)$ is plotted horizontally with the dashed line indicating the zero axis and positive values plotted to the left.

(E) The image is a static dark bar. The bar overlaps both an OFF (small τ) and an ON (large τ) region, generating opposing positive and negative contributions to $L(t)$.

(F) The weak response corresponding to E, plotted as in figure D.

The flashed dark bar of figures A-C is a more effective stimulus than the static bar of figure E.

Example 2.94 (Why a moving grating is a particularly effective stimulus)**.** The following figure shows responses to moving gratings estimated from a separable space-time receptive field. The receptive field is the same as in Example 2.93.



(A-C) The stimulus is a grating moving to the left.

- At the time corresponding to figure A, OFF regions overlap with dark bands and ON regions with light bands, generating a strong response.
- At the time of the estimate in figure B, the alignment is reversed, and $L(t)$ is negative.
- Figure C is a plot of $L(t)$ versus time corresponding to the responses generated in figures A-B. Time runs vertically in this plot and $L(t)$ is plotted horizontally, with the dashed line indicating the zero axis and positive values plotted to the left.

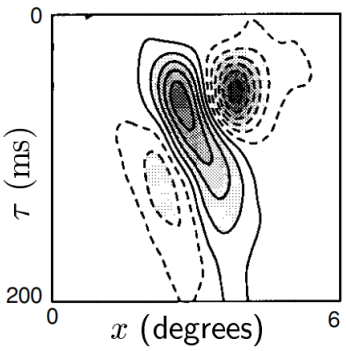
(D-F) The stimulus is a grating moving to the right. The responses are identical to those in figures A-C.

Remark 2.45. Separable space-time receptive fields can produce responses that are maximal for certain speeds of grating motion, but they are not sensitive to the direction of motion.

2.3.5 Nonseparable Receptive Fields

Remark 2.46. Many neurons in primary visual cortex are selective for the direction of motion of an image. Accounting for direction selectivity requires nonseparable space-time receptive fields.

Example 2.95. An example of a nonseparable receptive field is shown as below. This neuron has a three-lobed OFF-ON-OFF spatial receptive field, and these subregions shift to the left as time moves forward (and τ decreases). This means that the optimal stimulus for this neuron has light and dark areas that move toward the left.



Remark 2.47. One way to describe a nonseparable receptive field structure is to use a separable function constructed from a product of a Gabor function for D_s and Equation 2.41 for D_t, but to write these as functions of a mixture or rotation of the x and τ variables.

Lemma 2.96. The rotation matrix

$$M'(\psi) = \begin{pmatrix} \cos(\psi) & -\sin(\psi) \\ \sin(\psi) & \cos(\psi) \end{pmatrix} \tag{2.47}$$

can rotates a vector in two-dimensional space by ψ counterclockwise.

Proposition 2.97. The rotation of the space-time receptive field is achieved by mixing the space and time coordinates, using the transformation

$$D(x, y, \tau) = D_s(x', y) D_t(\tau') \tag{2.48}$$

with

$$x' = x\cos(\psi) - c\tau\sin(\psi) \tag{2.49}$$

and

$$\tau' = \tau\sin(\psi) + \frac{x}{c}\sin(\psi), \tag{2.50}$$

where factor c converts between the units of time (ms) and space (degrees), and ψ is the space-time rotation angle.

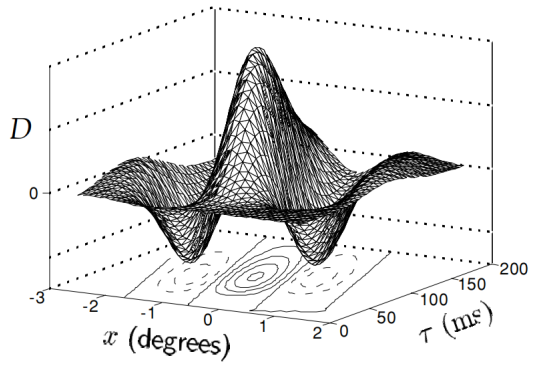
Proof. Note that the origin of x-τ plot in Example 2.95 is on the upper left corner, thus this rotation should be counterclockwise. Equation 2.49 and 2.50 follow from Lemma 2.96 directly. □

Example 2.98. Mathematical description of the space-time receptive field in Example 2.95 constructed from equations 2.48 - 2.50. The parameters are selected as follows:

(i) The Gabor function used (evaluated at $y = 0$) had $\sigma_x = 1°$, $1/k = 0.5°$, and $\phi = 0$.

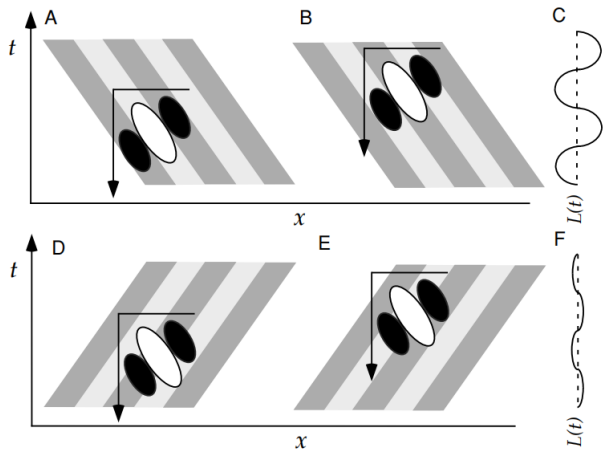
(ii) D_t is given by the expression in Equation 2.41 with $\alpha = 20$ ms, except that the second term, with the seventh power function, was omitted because the receptive field does not reverse sign in this example.

(iii) The x-τ rotation angle used was $\psi = \pi/9$, and the conversion factor was $c = 0.02°/\text{ms}$.



Remark 2.48. The rotation operation is not the only way to generate nonseparable space-time receptive fields. They are often constructed by adding together two or more separable space-time receptive fields with different spatial and temporal characteristics.

Example 2.99 (Direction Sensitivity)**.** The following figures show responses to moving gratings estimated from a nonseparable spacetime receptive field.



(A-C) The stimulus is a grating moving to the left.

- At the time corresponding to figure A, OFF regions overlap with dark bands and the ON region overlaps a light band, generating a strong response.
- At the time of the estimate in figure B, the alignment is reversed, and $L(t)$ is negative.
- Figure C is a plot of $L(t)$ versus time corresponding to the responses generated in figures A-B.

(D-F) The stimulus is a grating moving to the right. Because of the tilt of the space-time receptive field, the alignment with the right-moving grating is never optimal and the response is weak (figure F).

Remark 2.49. Although x-corrdination of the receptive field in the Example 2.99 changes over time, the variation range of x is only the range displayed by the three oval regions. x does not always move in one direction over time, but move periodically within this range over time.

Remark 2.50. As a result, a neuron with a nonseparable space-time receptive field can be *selective for the direction*

of motion of a grating and *for its velocity*, responding most vigorously to an optimally spaced grating moving at a velocity given by $c\tan(\psi)$. That is, the moving velocity of the space-time receptive field is the *preferred velocity*.

2.3.6 Static Nonlinearities: Simple Cells

Remark 2.51. Once the linear response estimate $L(t)$ has been computed, the firing rate of a visually responsive neuron can be approximated by using Equation 2.18, $r_{\text{est}}(t) = r_0 + F(L(t))$, where F is an appropriately chosen static nonlinearity.

Example 2.100. The simplest choice for F consistent with the positive nature of firing rates is rectification, $F = G[L]_+$, with G set to fit the magnitude of the measured firing rates.

Remark 2.52. However, the choice in Example 2.100 makes the firing rate a linear function of the contrast amplitude, which does not match the data on the contrast dependence of visual responses.

Definition 2.101. The *contrast saturation* means neural responses saturate as the contrast of the image increases, and are more accurately described contrast saturation by

$$r \propto A^n/(A_{1/2}^n + A^n)$$

where n is near 2, and $A_{1/2}$ is a parameter equal to the contrast amplitude that produces a half-maximal response.

Proposition 2.102. A static nonlinearity defined by

$$F(L) = \frac{G[L]_+^2}{A_{1/2} + G[L]_+^2} \tag{2.51}$$

reproduces the observed contrast dependence.

2.4 Static Nonlinearities: Complex Cells

Remark 2.53. The spatial receptive fields of complex cells cannot be divided into separate ON and OFF regions that sum linearly to generate the response. Areas where light and dark images excite the neuron overlap, making it difficult to measure and interpret spike-triggered average stimuli.

Principle 2.103. Like simple cells, complex cells are *selective to the spatial frequency and orientation* of a grating. However, unlike simple cells, complex cells respond to bars of light or dark no matter where they are placed within the overall receptive field. Likewise, the responses of complex cells to grating stimuli show *little dependence on spatial phase*.

Definition 2.104. The phenomenon that a complex cell is selective for a particular type of image independent of its exact spatial position within the receptive field is called the *spatial-phase invariance*.

Remark 2.54. The spatial-phase invariance of complex cells may represent an early stage in the visual processing that ultimately leads to position-invariant object recognition.

Remark 2.55. Complex cells also have temporal response characteristics that distinguish them from simple cells.

Principle 2.105. Complex cell responses to moving gratings are approximately constant, not oscillatory as in examples 2.94 and 2.99. The firing rate of a complex cell responding to a counterphase grating oscil lating with frequency ω has both a constant component and an oscillatory component with a frequency of 2ω, a phenomenon known as *frequency doubling*.

Remark 2.56. To give a first approximation to complex-cell responses, the key observation comes from Equation 2.46, which shows how the linear response estimate of a simple cell depends on spatial phase for an optimally oriented grating with K not too small.

Proposition 2.106. Consider two such responses, labeled L_1 and L_2, with preferred spatial phases ϕ and $\phi - \pi/2$. Including both the spatial and the temporal response factors, we find, for preferred spatial phase ϕ

$$L_1 = AB(\omega, K)\cos(\phi - \Phi)\cos(\omega t - \delta), \tag{2.52}$$

where $B(\omega, K)$ is a temporal and spatial frequency-dependent amplitude factor. For preferred spatial phase $\phi - \pi/2$,

$$L_2 = AB(\omega, K)\sin(\phi - \Phi)\cos(\omega t - \delta). \tag{2.53}$$

Thus we have,

$$L_1^2 + L_2^2 = A^2B^2(\omega, K)\cos^2(\omega t - \delta). \tag{2.54}$$

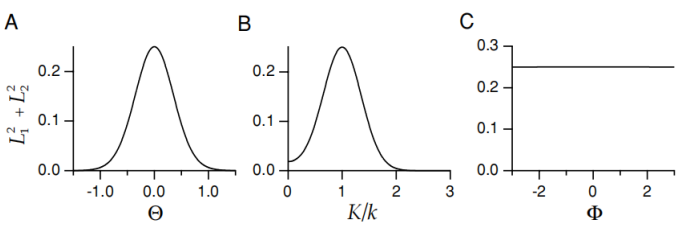
Proof. Equation 2.52 follows from Equation 2.46, Equation 2.53 follows from $\cos(\phi - \pi/2 - \Phi) = \sin(\phi - \Phi)$ and Equation 2.54 follows from $\cos^2(\phi - \Phi) + \sin^2(\phi - \Phi) = 1$. □

Corollary 2.107. We can describe the spatial-phase-invariant response of a complex cell by writing

$$r(t) = r_0 + G(L_1^2 + L_2^2), \tag{2.55}$$

for some constant G.

Example 2.108. Selectivity of the complex cell model (Equation 2.55) in response to a sinusoidal grating is shown in the following figures. The width and preferred spatial frequency of the Gabor functions underlying the estimated firing rate satisfy $k\sigma = 2$.



A The complex cell response estimate, $L_1^2 + L_2^2$, as a function of stimulus orientation Θ for a grating with the preferred spatial frequency $K = k$.

B $L_1^2 + L_2^2$ as a function of the ratio of the stimulus spatial frequency to its preferred value, K/k, for a grating oriented in the preferred direction $\Theta = 0$.

C $L_1^2 + L_2^2$ as a function of stimulus spatial phase Φ for a grating with the preferred spatial frequency and orientation, $K = k$ and $\Theta = 0$.

Remark 2.57. The response of the model complex cell is tuned to orientation and spatial frequency, but the spatial phase dependence, illustrated for a simple cell in figure C of Example 2.108, is absent. In computing the curve for figure C in Example 2.108, we used the exact expressions for L_1 and L_2 from the integrals in equations 2.43 and 2.44, not the approximation used in Equation 2.46 to simplify the previous discussion. Although it is not visible in the figure, there is a weak dependence on Φ when the exact expressions are used.

Proposition 2.109. The complex cell response given by equations 2.54 and 2.55 reproduces the frequency-doubling effect seen in complex cell responses.

Proof. This follows from the identity

$$\cos^2(\omega t - \delta) = \frac{1}{2}\cos(2(\omega t - \delta)) + \frac{1}{2}, \qquad (2.56)$$

where the last term on the right side of this equation generates the constant component of the complex cell response to a counterphase grating. □

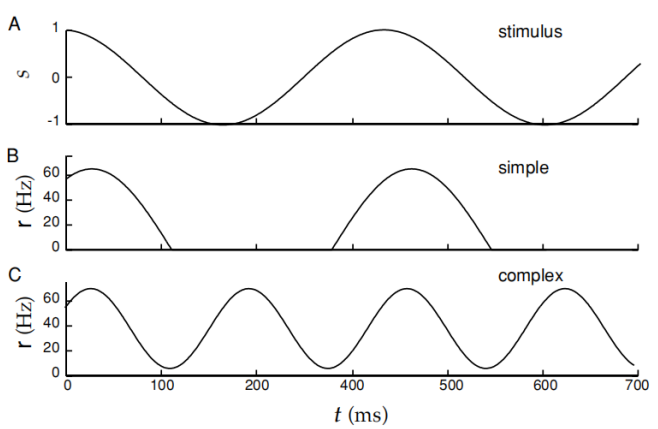
Example 2.110. Temporal responses of model simple and complex cells to a counterphase grating is shown in the following figures.

(A) The stimulus $s(x, y, t)$ at a given point (x, y) plotted as a function of time.

(B) The rectified linear response estimate of a model simple cell to this grating with a temporal kernel given by Equation 2.41 with $\alpha = 1/(15$ ms).

(C) The "frequency-doubled" response of a model complex cell with the same temporal kernel but with the estimated rate given by a squaring operation rather than rectification. The background firing rate is $r_0 = 5$ Hz.

Note the temporal phase shift of both B and C relative to A.



Definition 2.111. The description of a complex cell response that we have presented is called an *energy model* because of its resemblance to the equation for the energy of a simple harmonic oscillator. The pair of linear filters used, with preferred spatial phases separated by $\pi/2$, is called a *quadrature pair*.

Proposition 2.112. We can write the complex cell response as the sum of the squares of four rectified simple cell responses,

$$r(t) = r_0 + G([L_1]_+^2 + [L_2]_+^2 + [L_3]_+^2 + [L_4]_+^2), \qquad (2.57)$$

where the different $[L]_+$ terms represent the responses of simple cells with preferred spatial phases ϕ, $\phi + \pi/2$, $\phi + \pi$, and $\phi + 3\pi/2$.

Proof. Because of rectification, the terms L_1^2 and L_2^2 cannot be constructed by squaring the outputs of single simple cells. However, they can each be constructed by summing the squares of rectified outputs from two simple cells with preferred spatial phases separated by π. □

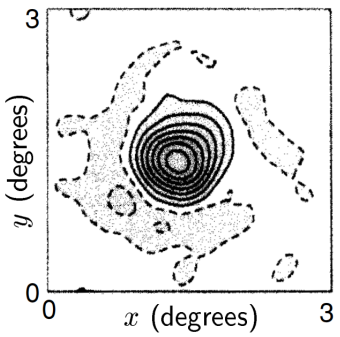
Remark 2.58. While such a construction is possible, it should not be interpreted too literally because complex cells receive input from many sources, including the LGN and other complex cells. Rather, this model should be viewed as purely descriptive. Simple mechanistic models of complex cells are described at the end of this chapter.

2.5 Receptive Fields in the Retina and LGN

Definition 2.113. A receptive field with a center-surround structure consisting either of a circular central ON region surrounded by an annular OFF region is called *ON-center*, or the opposite arrangement of a central OFF region surrounded by an ON region is called *OFF-center*.

Remark 2.59. Retinal ganglion cells display a wide variety of response characteristics, including nonlinear and direction-selective responses. However, a class of retinal ganglion cells (X cells in the cat or P cells in the monkey retina and LGN) can be described by a linear model built using reverse-correlation methods. The receptive fields of this class of retinal ganglion cells and an analogous type of LGN relay neurons are similar. The receptive fields of these neurons are ON-center or OFF-center.

Example 2.114. The following figure shows the center-surround spatial structure of the receptive field of a cat LGN X cell. This has a central ON region (solid contours) and a surrounding OFF region (dashed contours).

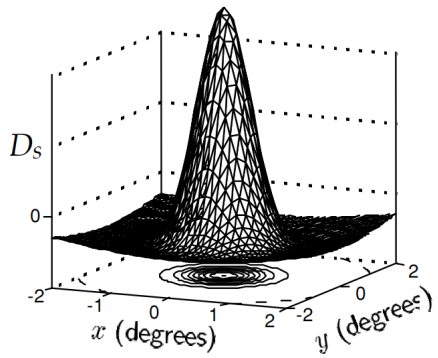


Definition 2.115. A *difference-of-Gaussians model* capturing the spatial structure of retinal ganglion and LGN receptive fields is expressed as

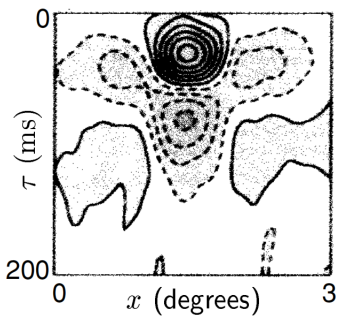
$$D_s(x,y) = \pm\left(\frac{1}{2\pi\sigma_{\text{cen}}^2}e^{-\frac{x^2+y^2}{2\sigma_{\text{cen}}^2}} - \frac{B}{2\pi\sigma_{\text{sur}}^2}e^{-\frac{x^2+y^2}{2\sigma_{\text{sur}}^2}}\right), \quad (2.58)$$

where the first Gaussian function describes the center, the second describes the surround, σ_{cen} determines the size of the central region, σ_{sur}, which is greater than σ_{cen}, determines the size of the surround, B controls the balance between center and surround contributions, the $\pm$ sign allows both ON-center (+) and OFF-center (−) cases to be represented.

Example 2.116. A fit of the receptive field shown in Example 2.114 using a difference-of-Gaussians function (Equation 2.58) with $\sigma_{\text{cen}} = 0.3°$, $\sigma_{\text{csur}} = 1.5°$, and $B = 5$.



Example 2.117. The following figure shows the space-time receptive field of a cat LGN X cell. Note that the center and surround regions both reverse sign as a function of τ and that the temporal evolution is slower for the surround than for the center.



Because of the difference between the time course of the center and of the surround regions, the space-time receptive field is not separable, although the center and surround components are individually separable.

Definition 2.118. A model capturing basic features of LGN neuron space-time receptive fields is expressed as

$$D(x,y,\tau) = \pm\left(\frac{D_t^{\text{cen}}(\tau)}{2\pi\sigma_{\text{cen}}^2}e^{-\frac{x^2+y^2}{2\sigma_{\text{cen}}^2}} - \frac{BD_t^{\text{sur}}(\tau)}{2\pi\sigma_{\text{sur}}^2}e^{-\frac{x^2+y^2}{2\sigma_{\text{sur}}^2}}\right), \quad (2.59)$$

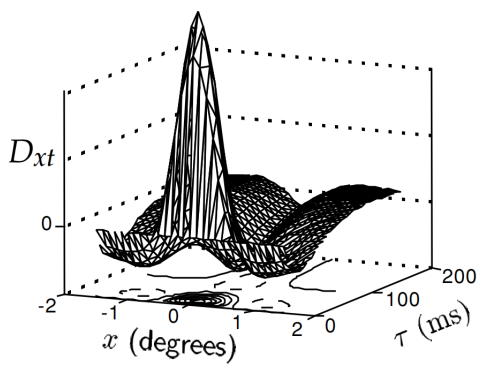
where $D_t^{\text{cen}}(\tau)$ and $D_t^{\text{sur}}(\tau)$ can both be described by the same functions, using two sets of parameters,

$$D_t^{\text{cen,sur}}(\tau) = \alpha_{\text{cen,sur}}^2\tau e^{-\alpha_{\text{cen,sur}}\tau} - \beta_{\text{cen,sur}}^2\tau e^{-\beta_{\text{cen,sur}}\tau}, \quad (2.60)$$

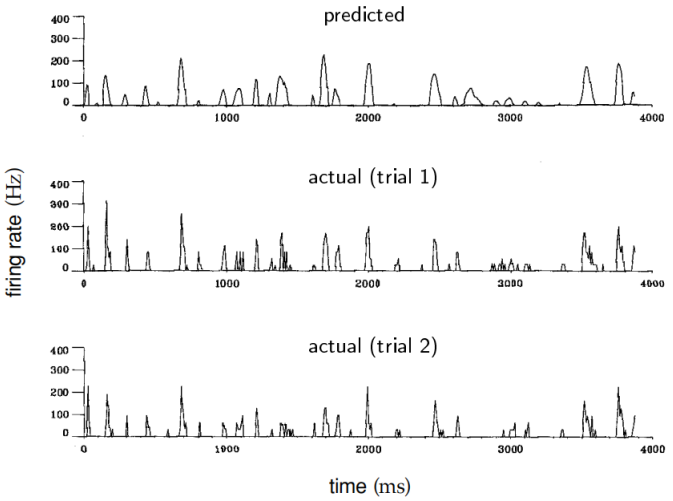
where α_{cen} and α_{sur} control the latency of the response in the center and surround regions, respectively, and β_{cen} and β_{sur} affect the time of the reversal.

Remark 2.60. The function in Equation 2.60 has characteristics similar to the function in Equation 2.41, but the latency effect is less pronounced.

Example 2.119. A fit of the space-time receptive field in Example 2.117 using Equation 2.59 with the same parameters for the Gaussian functions as in Example 2.116, and temporal factors given by Equation 2.60 with $1/\alpha_{\text{cen}} = 16$ ms, $1/\alpha_{\text{sur}} = 32$ ms, and $1/\beta_{\text{cen}} = 1/\beta_{\text{sur}} = 64$ ms.



Example 2.120 (A direct test of a reverse-correlation model of an LGN neuron)**.** Comparison of predicted and measured firing rates for a cat LGN neuron responding to a video movie is shown in the following figures.



(i) The top panel is the rate predicted by integrating the product of the video image intensity and the kernel needed to describe this neuron was first extracted by using a white-noise stimulus. The resulting linear prediction was rectified, that is, $F(L) = G[L]_+$.

(ii) The middle and lower panels are measured firing rates extracted from two different sets of trials.

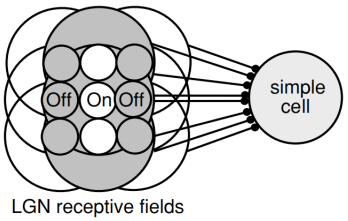
Remark 2.61. In Example 2.120, the correlation coefficient between the predicted and actual firing rates was 0.5, which was very close to the correlation coefficient between firing rates extracted from different groups of trials. This means that the error of the prediction was no worse than the variability of the neural response itself.

2.6 V1 Receptive Fields Construction

Remark 2.62. The models of visual receptive fields we have been discussing are purely descriptive, but they provide an important framework for studying how the circuits of the retina, LGN, and primary visual cortex generate neural responses.

Definition 2.121. The *Hubel-Wiesel Model* propose the oriented receptive fields of cortical neurons could be generated by summing the input from appropriately selected LGN neurons.

Example 2.122 (The Hubel-Wiesel model of a simple cell)**.** The Hubel-Wiesel model of orientation selectivity is shown below.



The spatial arrangement of the receptive fields of nine LGN neurons are shown, with a column of three ON-center fields flanked on either side by columns of three OFF-center fields. White areas denote ON fields and gray areas, OFF fields. In the model, the converging LGN inputs are summed by the simple cell. *This arrangement produces a receptive field oriented in the vertical direction.* Note that, two center types are represented as

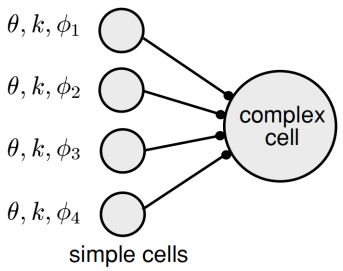
(i) ON-center field: a concentric circle with an white inner circle and an gray outer circle;

(ii) OFF-center field: a concentric circle with an gray inner circle and an white outer circle.

Remark 2.63. This model accounts for the selectivity of a simple cell purely on the basis of feedforward input from the LGN.

Remark 2.64. In a previous section, we showed how the properties of complex cell responses could be accounted for by using a squaring static nonlinearity. While this provides a good description of complex cells, there is little indication that complex cells actually square their inputs. Models of complex cells can be constructed without introducing a squaring nonlinearity.

Example 2.123 (The Hubel-Wiesel model of a complex cell)**.** Inputs from a number of simple cells with similar orientation and spatial frequency preferences (θ and k), but different spatial phase preferences (ϕ_1, ϕ_2, ϕ_3, and ϕ_4), converge on a complex cell and are summed. *This produces a complex cell output that is selective for orientation and spatial frequency, but not for spatial phase (phase-invariant response).*



The figure shows four simple cells converging on a complex cell, but additional simple cells can be included to give a more complete coverage of spatial phase.

Remark 2.65. While the model generates complex cell responses, there are indications that complex cells in primary visual cortex are not driven exclusively by simple cell input. An alternative model is considered in chapter 7.

2.7 Questions

This section states the questions that we can't solve or the concepts that we can't understand.

2.7.1 the Bandwidth

This subsection belongs to Chapter 2 section 2.3.
When referring to the number of sub-regions of receptive region, a concept-bandwidth is introduced, but we don't know why it is introduced and what its geometric meaning is. The relevant key points are as follows.

Remark 2.66. The number of subregions within the receptive field is determined by the product $k\sigma_x$ and is typically expressed in terms of a quantity known as the bandwidth b.

Definition 2.124. The *bandwidth* is the width of the spatial frequency tuning curve measured in octaves (**we can't understand this word**) and defined as

$$b = \log_2(K_+/K_-),$$

where $K_+ > k$ and $K_- < k$ are the spatial frequencies of gratings that produce one-half the response amplitude of a grating with $K = k$.

Proposition 2.125. The relationship between $k\sigma_x$ and the bandwidth b is

$$b = \log_2\left(\frac{k\sigma_x + \sqrt{2\ln(2)}}{k\sigma_x - \sqrt{2\ln(2)}}\right) \text{ or } k\sigma_x = \sqrt{2\ln(2)}\frac{2^b+1}{2^b-1}. \tag{2.61}$$

Proof. The spatial frequency tuning curve as a function of K for a Gabor receptive field with preferred spatial frequency k and receptive field width σ_x is proportional to $\exp(-\sigma_x^2(k-K)^2/2)$ (see Equation 2.46). The values of K_+ and K_- needed to compute the bandwidth are thus determined by the condition $\exp(-\sigma_x^2(k-K_\pm)^2/2) = 1/2$. Solving this equation gives $K_\pm = k \pm \sqrt{2\ln(2)}/\sigma_x$, from which we obtain Equation 2.61. □

Remark 2.67. Bandwidth is defined only if $k\sigma_x > \sqrt{2\ln(2)}$, but this is usually the case. Bandwidths typically range from about 0.5 to 2.5, corresponding to $k\sigma_x$ between 1.7 and 6.9.

Remark 2.68. High bandwidths correspond to low values of $k\sigma_x$, meaning that the receptive field has few subregions and poor spatial frequency selectivity. Neurons with more subfields are more selective to spatial frequency, and they have smaller bandwidths and larger values of $k\sigma_x$.

Solution. The bandwidth means the interval range of x that can reach more than half of the extreme value.

Chapter 3

Neural Decoding

3.1 Encoding and Decoding

Notation 14. Several different probabilities and conditional probabilities enter into our discussion. The probabilities we need are:

- $P[s]$, the probability of stimulus s being presented, often called the prior probability.
- $P[\mathbf{r}]$, the probability of response $\mathbf{r}$ being recorded independent of what stimulus was used.
- $P[\mathbf{r}, s]$, the probability of stimulus s being presented and response $\mathbf{r}$ being recorded. This is called the joint probability.
- $P[\mathbf{r}|s]$, the conditional probability of evoking response $\mathbf{r}$, given that stimulus s was presented.
- $P[s|\mathbf{r}]$, the conditional probability that stimulus s was presented, given that response $\mathbf{r}$ was recorded.

Here $\mathbf{r} = (r_1, r_2, \cdots, r_N)$ for N neurons is a list of spike-count fring rates.

Notation 15. The *encoding* is characterized by the set of probabilities $P[\mathbf{r}|s]$ for all stimuli and responses, with which we considered the problem of predicting neural responses to known stimuli. The *decoding* a response is to determine the probabilities $P[s|\mathbf{r}]$ which reflects what is going on in the real world from neuronal spiking patterns.

Theorem 3.1. *Bayes theorem* relating $P[s|\mathbf{r}]$ to $P[s|\mathbf{r}]$:

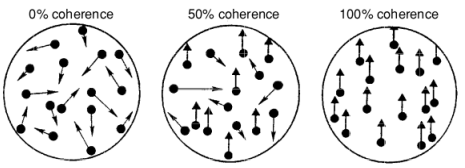
$$P[s|\mathbf{r}] = \frac{P[\mathbf{r}|s]P[s]}{P[\mathbf{r}]}.$$

Remark 3.1. According to Bayes theorem, $P[s|\mathbf{r}]$ can be obtained from $P[\mathbf{r}|s]$, but the stimulus probability $P[s]$ is also needed. As a result, decoding requires knowledge of the statistical properties of experimentally or naturally occurring stimuli.

Remark 3.2. We sometimes treat the response fring rates or the stimulus values as continuous variables. In this case, the probabilities listed must be replaced by the corresponding probability densities, $p[\mathbf{r}]$, $p[\mathbf{r}|s]$, etc.

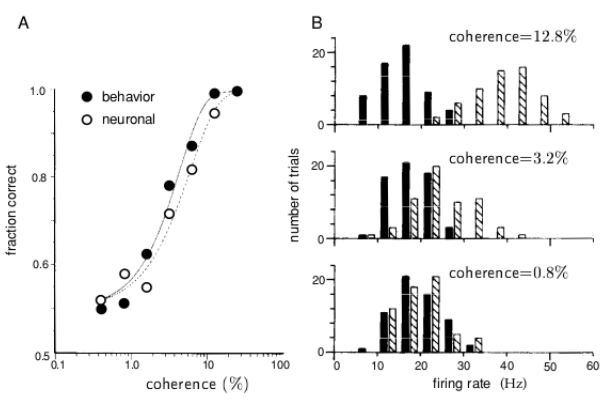
3.2 Discrimination

Example 3.2. In the experiments performed by Britten et al. (1992)., a monkey was trained to discriminate between two directions of motion of a visual stimulus which was a pattern of dots on a video monitor. The percentage of dots that move together in the fixed direction is called the coherence level. By varying the degree of coherence shown by pictures, the task of detecting the movement direction can be made more or less diffcult.



Definition 3.3 (plus and minus)**.** The preferred direction was called *plus* (or +) direction that produced the maximum response in that neuron, and its opposite direction is called the *minus* (or −) direction.

Example 3.4. During the same experiment in Example 3.2, the judgment accuracy of the monkey and the optic nerve coding signal activity in the MT area were recorded. The experimental results show that: first, the coding of MT neural activity is basically sufficient for judging the direction; second, at high coherence levels, the fring-rate distributions corresponding to the two directions are fairly well separated, while at low coherence levels, they merge.



Remark 3.3. Although spike count rates take only discrete values, it is more convenient to treat r as a continuous variable for our discussion. Treated as probability densities, these two distributions are approximately Gaussian with the same variance, σ_r^2, but different means, $\langle r \rangle_+$ for the plus direction and $\langle r \rangle_-$ for the minus direction.

Definition 3.5 (discriminability)**.** A convenient measure of the separation between the distributions is the *discriminability*

$$d' = \frac{\langle r \rangle_+ - \langle r \rangle_-}{\sigma_r}. \tag{3.1}$$

Remark 3.4. Decoding involves using the neural response to determine in which of the two possible directions the stimulus moved for Example 3.2. A simple decoding procedure in this case is to determine the fring rate r during a trial and compare it to a threshold number z. If $r \geq z$, we report "plus"; otherwise we report "minus".

Definition 3.6 (size and power)**.** Below are the probabilities of answering plus for both given the conditions:

(i) The probability that it will give the answer "plus" when the stimulus is moving in the plus direction is the conditional probability that $r \geq z$ given a plus stimulus, $\alpha(z) = P[r \geq z|+]$, called *size* or *false alarm* rate of the test.

(ii) 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 $\beta(z) = P[r \geq z|-]$, called *power* or *hit* rate of the test.

These two probabilities completely determine the performance of the decoding procedure because the probabilities for the other two cases

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

3.2.1 ROC Curves

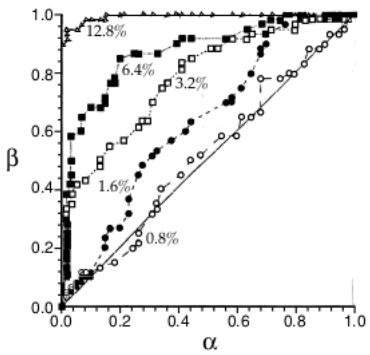
Definition 3.7. The *receiver operating characteristic* (**ROC**) curve is traced out as a function of the threshold z. Each point on an ROC curve corresponds to a different value of z. The x coordinate of the point is α, the size of the test for this value of z and the y coordinate is β. ROC curve provides a way of evaluating how test performance depends on the choice of z and indicates how the size and power of a test trade off as the threshold is varied.

Example 3.8. The figure shows ROC curves computed by Britten et al. for several different values of the stimulus coherence.

(i) At high coherence levels, when the task is easy, the ROC curve rises rapidly from $\alpha(z) = 0$, $\beta(z) = 0$ as the threshold is lowered from a high value, and the probability $\beta(z)$ of a correct "plus" answer quickly approaches 1 without a concomitant increase in $\alpha(z)$. 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.

(ii) At lower high coherence levels, when the task is difficult, the curve rises more slowly as z is lowered.

(iii) At quite low coherence levels, the task is impossible, in that the test merely gives random answers, the curve will lie along the diagonal $\alpha = \beta$, because the probabilities of answers being correct and incorrect are equal.



Remark 3.5. Examination of Example 3.8 suggests a relationship between the area under the ROC curve and the level of performance on the task. When the ROC curve in Example 3.8 lies along the diagonal, the area underneath it is 1/2, which is the probability of a correct answer in this case (given any threshold). When the task is easy and the ROC curve hugs the left axis and upper limit, the area under it approaches 1, which is again the probability of a correct answer (given an appropriate threshold). The area underneath the ROC curve is the probability of a correct answer in the most cases (given an appropriate threshold).

However, the precise relationship between task performance and the area under the ROC curve is complicated by the fact that different threshold values can be used. This ambiguity can be removed by considering a slightly different task, called *two-alternative forced choice*.

Notation 16. For *two-alternative forced choice*, the stimulus is presented twice, once with motion in the plus direction and once in the minus di- rection. The task is to decide which presentation corresponded to the plus direction, given the fring 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 threshold variable from consideration.

Proposition 3.9. In the two-alternative force-choice task, the value of r on one trial serves as the threshold for the other trial. Then the probability of getting the correct answer

$$P[\text{correct}] = \int_0^\infty \text{p[z|}-\text{]}\beta(\text{z})\text{dz}. \tag{3.2}$$

Proof. 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 $z = r_2$.

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 fring-rate probability density for a fring rate $r = z$. Integrating over all values of z gives the answer. □

Proposition 3.10. The probability of getting the correct answer in the Equation 3.2 can be transformed into

$$P[\text{correct}] = \int_0^1 \beta \mathrm{d}\alpha. \tag{3.3}$$

Proof. $\alpha(z)$ mentioned in definition 3.6, can be written as an integral of the conditional fring-rate probability density $p[r|-]$,

$$\alpha(z) = \int_z^\infty p[r|-]dr. \tag{3.4}$$

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

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

This allows us to make the replacement $p[z|-]dz \to -d\alpha$ in the integral of Equation (3.2) and to change the integration variable from z to α. Noting that $\alpha = 1$ when $z = 0$ and $\alpha = 0$ when $z = \infty$, we infer it. □

Remark 3.6. The ROC curve is just β plotted as a function of α, so this integral is the area under the ROC curve. Thus, the area under the ROC curve is the probability of responding correctly in the two-alternative forced-choice test.

Exercise 3.11. Prove that suppose that $p[r|+]$ and $p[r|-]$ are both Gaussian functions with means $\langle r\rangle_+$ and $\langle r\rangle_-$, and a common variance σ_r^2. The reader is invited to show that, in this case,

$$P[\text{correct}] = \frac{1}{2}\text{erfc}\Big(\frac{\langle r\rangle_+ - \langle r\rangle_-}{2\sigma_r}\Big) = \frac{1}{2}\text{erfc}\Big(-\frac{d'}{2}\Big), \tag{3.5}$$

where d' is the discriminability defined in equation (3.1) and erfc(x) is the complementary error function (which is an integral of a Gaussian distribution) defined as

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

Remark 3.7. $P[\text{correct}]$ and d' are positively correlated, that is to say, the greater the difference in their firing rates, the greater the probability of accurate judgment. And in the case where the distributions are equalvariance Gaussians, the relationship between the discriminability and the area under the ROC curve is invertible because the complementary error function is monotonic.

3.2.2 ROC Analysis of Motion Discrimination

Remark 3.8. To interpret the experiment as a two-alternative forced-choice task, Brit ten et al. imagined that, in addition to being given the fring rate of the recorded neuron during stimulus presentation, the observer is given the fring rate of a hypothetical "anti-neuron" having response characteristics exactly opposite from the recorded neuron. In reality, the responses of this anti-neuron 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 will also reappear in our discussion of spike-train decoding. An observer predicting motion directions on the basis of just these two neurons at a level equal to the area under the ROC curve is termed an ideal observer.

Remark 3.9. The figure A in Example 3.4 shows a typical result for the performance of an ideal observer using one recorded neuron and its anti-neuron partner. The open circles in figure were obtained by calculating the areas under the ROC curves for this neuron. Amazingly, the ability of the ideal observer to perform the discrimination task using a single neuron/anti-neuron 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 neurons, while the ideal observer uses only two.

3.2.3 The Likelihood Ratio Test

Lemma 3.12. The discrimination test we have considered compares the fring rate to a fixed threshold value. The Neyman-Pearson lemma shows that it is optimal to choose the test function the ratio of probability densities (or probabilities),which also can be seen function of the fring rate

$$l(r) = \frac{p[r|+]}{p[r|-]}, \tag{3.6}$$

which is known as the *likelihood ratio.*

Proof. Consider the difference β 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 $h(r)$. For the test $h(r)$ using the threshold z_h,

$$\begin{aligned}\alpha_h(z_h) &= \int p[r|-]\Theta(h(r) - z_h)dr,\\ \beta_h(z_h) &= \int p[r|+]\Theta(h(r) - z_h)dr.\end{aligned} \tag{3.7}$$

Similar equations hold for the $\alpha_l(z_l)$ and $\beta_l(z_l)$ values for the test $l(r)$ using the threshold z_l. We use the Θ function, which is 1 for positive and 0 for negative values of its argument, to impose the condition that the test is greater than the threshold. Comparing the β values for the two tests, we find

$$\begin{aligned}\nabla\beta &= \beta_l(z_l) - \beta_h(z_h)\\ &= \int p[r|+]\Theta(l(r) - z_l)dr - intp[r|+]\Theta(h(r) - z_h)dr.\end{aligned} \tag{3.8}$$

The range of integration where $l(r) \geq z_l$ and also $h(r) \geq z_h$ cancels between these two integrals and use the definition $l(r) = p[r|+]/p[r|-]$, we can replace $p[r|+]$ with $l(r)p[r|-]$

in this equation, giving

$$\nabla\beta = \int l(r)p[r|-]\left(\Theta(l(r)-z_l)\Theta(z_h-h(r))\right)dr - \int l(r)p[r|-]\left(\Theta(z_l-l(r))\Theta(h(r)-z_h)\right)dr. \tag{3.9}$$

Then, due to the conditions imposed on $l(r)$ by the Θ functions within the integrals, replacing $l(r)$ by z can neither decrease the value of the integral resulting from the first term in the large parentheses, nor increase the value arising from the second. This leads to the inequality

$$\nabla\beta \geq z\int p[r|-]\Theta(l(r)-z_l)\Theta(z_h-h(r))dr - z\int p[r|-]\Theta(z_l-l(r))\Theta(h(r)-z_h)dr. \tag{3.10}$$

Putting back the region of integration that cancels between these two terms (for which $l(r) \geq z_l$ and $h(r) \geq z_h$), we find

$$\nabla\beta \geq z\Big[\int p[r|-]\Theta(l(r)-z_l)dr - \int p[r|-]\Theta(h(r)-z_h)dr\Big]. \tag{3.11}$$

By definition, these integrals are the sizes of the two tests, which are equal by hypothesis. Thus $\beta \geq 0$, showing likelihood ratio $l(r)$, at least in the sense of maximizing the power for a given size. □

Remark 3.10. The test function r used above is not equal to the likelihood ratio. However, if the likelihood is a monotonically increasing function of r, the fring-rate threshold test is equivalent to using the likelihood ratio and is also 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.

Proposition 3.13. There is a direct relationship between the likelihood ratio and the ROC curve. As in Equation 3.4 and (3.10), we can write

$$\beta(z) = \int_z^\infty p[r|+]dr \quad \text{so} \quad \frac{d\beta}{dz} = -p[z|+]. \tag{3.12}$$

Combining this result with Equation 3.10, we find that

$$\frac{d\beta}{d\alpha} = \frac{d\beta}{dz}\frac{dz}{d\alpha} = l(z), \tag{3.13}$$

so the slope of the ROC curve is equal to the likelihood ratio.

Remark 3.11. 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.

Definition 3.14. The penalty associated with answering "minus" when the correct answer is "plus" is quantifed by the *loss parameter* L_-. Similarly, quantify the loss for answering "plus" when the correct answer is "minus" as L_+.

Theorem 3.15. The probabilities that the correct answer is "plus" or "minus", given the fring rate r, are $P[+|r]$ and $P[-|r]$ respectively. These probabilities are related to the conditional fring-rate probability densities by Bayes theorem,

$$P[+|r] = \frac{p[r|+]P[+]}{p[r]} \quad \text{and} \quad P[-|r] = \frac{p[r|-]P[-]}{p[r]}. \tag{3.14}$$

Proposition 3.16. The average loss expected for a "plus" answer when the fring rate is r is the loss associated with being wrong times the probability of being wrong, $\text{Loss}_+ = \text{L}_+\text{P}[-|\text{r}]$. Similarly, the expected loss when answering "minus" is $\text{Loss}_- = \text{L}_-\text{P}[+|\text{r}]$. A reasonable strategy is to cut the losses, answering "plus" if $\text{Loss}_+ \leq \text{Loss}_-$ and "minus" otherwise. Using Equation 3.14, we find that this strategy gives the response "plus" if

$$l(r) = \frac{p[r|+]}{p[r|-]} \geq \frac{L_+P[-]}{L_-P[+]}. \tag{3.15}$$

This shows that the strategy of comparing the likelihood ratio to a threshold is a way of minimizing the expected loss.

Example 3.17. If the conditional probability densities $p[r|+]$ and $p[r|-]$ are Gaussians with means r_+ and r_- and identical variances σ_r^2 , and $P[+] = P[-] = 1/2$, the probability $P[+|r]$ is a sigmoidal function of r,

$$P[+|r] = \frac{1}{1+\exp(-d'(r-r_{ave})/\sigma_r)}, \tag{3.16}$$

where $r_{\text{ave}} = (r_+ + r_-)/2$.

Example 3.18. We have thus far considered discriminating between two quite distinct stimulus values, plus and minus. Often we are interested in discriminating between two stimulus values $s + \nabla s$ and s that are very close to one another. In this case, the likelihood ratio is

$$\frac{p[r|s+\nabla s]}{p[r|s]} \approx \frac{p[r|s]+\nabla s\partial p[r|s]/\partial s}{p[r|s]} = 1 + \nabla\frac{\partial \ln p[r|s]}{\partial s}. \tag{3.17}$$

For small ∇s, a test that compares

$$Z(r) = \frac{\partial \ln p[r|s]}{\partial s}, \tag{3.18}$$

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

3.3 Population Decoding

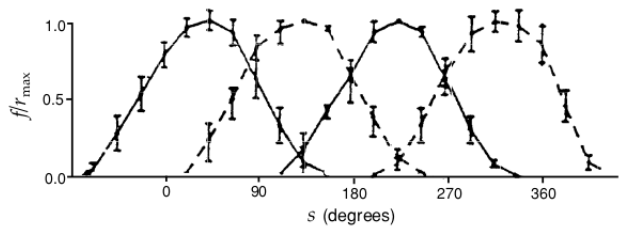
Remark 3.12. The use of large numbers of neurons to represent information is a basic operating principles of many nervous systems. *Population coding* has a number of advantages, including reduction of uncertainty due to neuronal variability and the ability to represent a number of different stimulus attributes simultaneously. In the previous section, we discussed discrimination between stimuli on the basis of the response of a single neuron. The responses of a population of neurons can also be used for discrimination, with

the only essential difference being 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 Neyman-Pearson lemma continue to apply in exactly the same way.

Remark 3.13. *Discrimination* is a special case of decoding in which only a few different stimulus values are considered. A more general problem is the extraction of a continuous stimulus parameter from one or more neuronal responses. In this section, we study how the value of a continuous parameter associated with a static stimulus can be decoded from the spike-count firing rates of a population of neurons.

3.3.1 Encoding and Decoding Direction

Example 3.19. The cercal system of the cricket, which senses the direction of incoming air currents is an interesting example of population coding involving a four interneurons. The figure shows average firing-rate tuning curves for the four relevant interneurons as a function of wind direction, which are well approximated by halfwave rectifed cosine function. The preferred directions of the neurons are located 90° from each other, and $r_{\max}$ values are typically around 40 Hz.



Proposition 3.20. 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 in response to wind direction s, $\langle r_a \rangle = f_a(s)$, normalized to its maximum, can be written as

$$\left(\frac{f(s)}{r_{\max}}\right)_a = [(\cos(s - s_a))]_+, \tag{3.19}$$

where the half-wave rectifcation eliminates negative firing rates. Here $r_{\max}$, which may be different for each neuron, is a constant equal to the maximum average firing rate.

Proposition 3.21. For the cercal system of the cricket, Equation 3.19 can be written as

$$\left(\frac{f(s)}{r_{\max}}\right)_a = [\vec{v} \cdot \vec{c}_a]. \tag{3.20}$$

where spatial vector $\vec{v}$ pointing parallel to the wind velocity in place of the angle s and having unit length $\left|\vec{v}\right| = 1$ and a vector $\vec{c}_a$ of unit length is the preferred wind direction pointing in the direction specified by the angle s_a for each interneuron.

Proof. In this case, we can use the vector dot product to write $\cos(s - s_a) = \vec{v} \cdot \vec{c}_a$. In terms of these vectors, the average firing rate is proportional to a half-wave rectified projection of the wind direction vector onto the preferred direction axis of the neuron. And combined with the above proposition 3.20, we give the answer. □

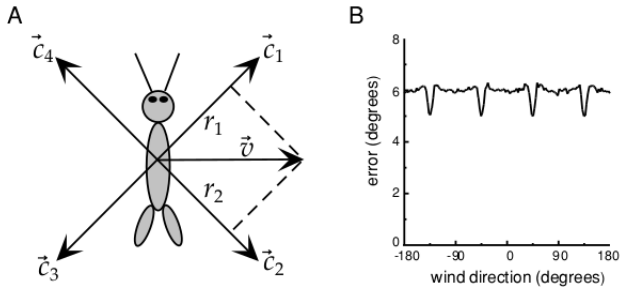
Remark 3.14. Decoding the cercal system is particularly easy because of the close relationship between the representation of wind direction it provides and a two-dimensional Cartesian coordinate system. The preferred directions of the four interneurons, like the x and y axes of a Cartesian coordinate system, lie along two perpendicular directions. Four neurons are required, rather than two, because firing rates cannot represent negative projections.

Proposition 3.22. If r_a is the spike-count fring rate of neuron a, an estimate of the wind direction on any given trial can be obtained from the direction of the vector

$$\vec{v}_{\text{pop}} = \sum_{a=1}^{4} \left(\frac{r}{r_{\max}}\right)_a \vec{c}_a. \tag{3.21}$$

This vector is known as the *population vector*, and the associated decoding method is called the *vector method*. In fact, this encoding is the one that requires the least number of neurons to encode two-dimensional directions.

Example 3.23. In figure A, preferred directions of four cercal interneurons in relation to the cricket's body. The 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. In figure B, the root-mean-square error in the wind direction determined by vector decoding of the firing rates of four cercal interneurons. These results 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 3.21 and compared to the wind direction used to generate them.



Example 3.24. 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, r_0. When an arm movement is made in the direction represented by a vector of

unit length, $\vec{v}$, the average fring rates for such a M1 neuron, labeled by an index a, can be written as

$$\left(\frac{\langle r\rangle - r_0}{r_{\max}}\right)_a = \left(\frac{f(s) - r_0}{r_{\max}}\right)_a = \vec{v}\cdot\vec{c}_a, \tag{3.22}$$

where $\vec{c}_a$ is the preferred-direction vector that defines the selectivity of the neuron. Unlike the cercal interneurons, M1 neurons do not have orthogonal preferred directions that form a Cartesian coordinate system. Instead, the preferred directions of the neurons appear to point in all directions with roughly equal probability.

Proposition 3.25. If the preferred directions point uniformly in all directions and the number of neurons N is suffciently large, the population vector

$$\vec{v}_{\text{pop}} = \sum_{a=1}^{N}\left(\frac{r - r_0}{r_{\max}}\right)_a \vec{c}_a, \tag{3.23}$$

will, on average, point in a direction parallel to the arm movement direction vector $\vec{c}$. If we average Equation 3.23 over trials and use Equation 3.22, we find

$$\langle\vec{v}_{\text{pop}}\rangle = \sum_{a=1}^{N}(\vec{v}\cdot\vec{c}_a)\vec{c}_a, \tag{3.24}$$

where $\vec{v}_{\text{pop}}$ 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.

3.3.2 Optimal Decoding Methods

Remark 3.15. The vector method is a simple decoding method that can perform quite well in certain cases, but it is neither a general nor an optimal way to reconstruct a stimulus from the fring rates of a population of neurons. In this section, we dicuss two methods, that are Bayesian inference and MAP inference, which are considered optimal by some measure.

Remark 3.16. *Bayesian* and *MAP* estimates use the conditional probability that a stimulus parameter takes a value between s and $s + \nabla s$, given that the set of N encoding neurons fired at rates given by $\mathbf{r}$. 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,

$$p[s|\mathbf{r}] = \frac{p[\mathbf{r}|s]p[s]}{p[\mathbf{r}]}. \tag{3.25}$$

A disadvantage of these methods is that extracting $p[s|\mathbf{r}]$ from experimental data can be diffcult.

Definition 3.26. Below are the two optimal ways to reconstruct a stimulus from the fring rates of a population of neurons:

(i) The *Bayesian inference* involves finding the minimum of a loss function $L(s, s_{\text{bayes}})$ that quantifyes the cost of reporting the estimate s_{bayes} when the correct answer is s.

(ii) The *MAP inference* chooses the stimulus value, s_{MAP}, that maximizes the conditional probability density of the stimulus, $p[s_{\text{MAP}}|\mathbf{r}]$ and *ML inference* chooses s_{ML} to maximize the likelihood function, $p[\mathbf{r}|s_{\text{ML}}]$, which generally produce estimates that are as accurate, in terms of the variance of the estimate, as any that can be achieved by a wide class of estimation methods (so-called unbiased estimates).

Remark 3.17. The value of s_{bayes} is chosen to minimize the expected loss averaged over all stimuli for a given set of rates, that is, to minimize the function $\int L(s, s_{\text{bayes}})p[s|\mathbf{r}]ds$.

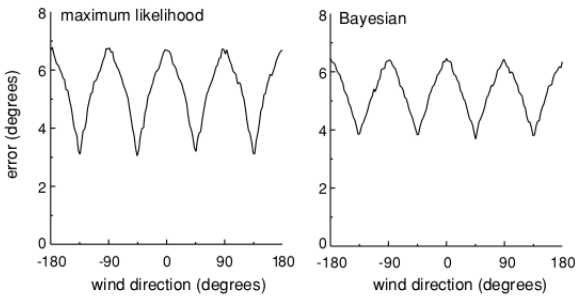
Example 3.27. If the loss function is the squared difference between the estimate and the true value, $L(s, s_{\text{bayes}}) = (s - s_{\text{bayes}})^2$, the estimate that minimizes the expected loss is the mean

$$s_{\text{bayes}} = \int p[s|\mathbf{r}]ds. \tag{3.26}$$

Example 3.28. If the loss function is the absolute value of the difference, $L(s, s_{\text{bayes}}) = \left|s - s_{\text{bayes}}\right|$, then s_{bayes} is the median rather than the mean of the distribution $p[s|\mathbf{r}]$.

Remark 3.18. The MAP approach is thus to choose as the estimate s_{MAP} the most likely stimulus value for a given set of rates. If the prior or stimulus probability density $p[s]$ is independent of s, then $p[s|\mathbf{r}]$ and $p[\mathbf{r}|s]$ have the same dependence on s, because the factor $p[s]/p[\mathbf{r}]$ in Equation 3.25 is independent of s. In this case, the MAP algorithm is equivalent to maximum likelihood (ML) inference.

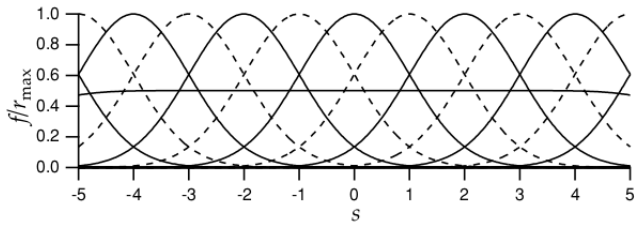
Example 3.29. The root-mean-squared difference between the true and estimated wind directions for the cercal system, using ML and Bayesian methods is shown as follows. The Bayesian estimate in figure is based on the squared-difference loss function. Both estimates use a constant stimulus probability density $p[s]$, so the ML and MAP estimates are identical. The Bayesian result has a slightly smaller average error across all angles. The dips in the error curves in figure appear at angles where one tuning curve peaks and two others rise from threshold. These dips are due to the two neurons responding near threshold, not to the maximally responding neu- ron. They occur because neurons are most sensitive at points where their tuning curves have maximum slopes, which in this case is near threshold.



Example 3.30. Up to now, we have considered the decoding of a direction angle. We now turn to the more general case of decoding an arbitrary continuous stimulus parameter. An instructive example is provided by an array of n neurons with preferred stimulus values distributed uniformly across the full range of possible stimulus values. An example of such an array for Gaussian tuning curves,

$$f_a(s) = r_{\max} \exp\left(-\frac{1}{2}\left(\frac{s - s_a}{\sigma_a}\right)^2\right). \tag{3.27}$$

In this example, each neuron has a tuning curve with a different preferred value s_a and potentially a different width σ_a. If the tuning curves are evenly and densely distributed across the range of s values, the sum of all tuning curves $\sum f_a(s)$ is approximately independent of s. The roughly flat line is proportional to this sum.



Remark 3.19. To implement the Bayesian, MAP, or ML approach, we need to know the conditional firing-rate probability density $p[\mathbf{r}|s]$ that describes this variability.

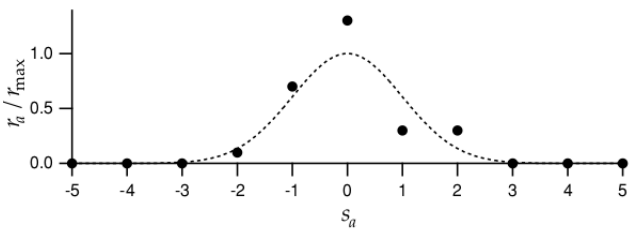
Proposition 3.31. We assume that the firing rate r_a of neuron a is determined by counting n_a spikes over a trial of duration T (so that $r_a = n_a/T$), and that the variability can be described by the homogeneous Poisson model discussed in chapter 1. In this case, the probability of stimulus s evoking $n_a = r_a T$ spikes, when the average firing rate is $\langle r_a \rangle = f_a(s)$, is given by

$$P[r_a|s] = \frac{(f_a(s)T)^{r_a T}}{(r_a T)!} \exp(-f_a(s)T). \tag{3.28}$$

If we assume that each neuron fires independently, the firing-rate probability for the population is the product of the individual probabilities,

$$P[\mathbf{r}|s] = \prod_{a=1}^{N} \frac{(f_a(s)T)^{r_a T}}{(r_a T)!} \exp(-f_a(s)T). \tag{3.29}$$

Example 3.32. The filled circles in figure show a set of randomly generated firing rates for the array of Gaussian tuning curves for $s = 0$ shown above. This figure also illustrates a useful way of visualizing population responses: plotting the responses as a function of the preferred stimulus values. The dashed curve is the tuning curve for the neuron with $s_a = 0$. Because the tuning curves are functions of $|s - s_a|$, the values of the dashed curve at $s_a = -5, -4, \ldots, 5$ are the mean activities of the cells with preferred values at those locations for a stimulus at $s = 0$.



Proposition 3.33. The ML estimated stimulus, s_{ML}, is the stimulus that maximizes $P[\mathbf{r}|s]$. We find that s_{ML} is determined by

$$T \sum_{a=1}^{N} r_a \frac{f_a'(s_{ML})}{f_a(s_{ML})} = 0,$$

where the prime denotes a derivative.

Proof. To apply the ML estimation algorithm, we only need to consider the terms in $P[\mathbf{r}|s]$ that depend on s. Because Equation 3.29 involves a product, it is convenient to take its logarithm and write

$$\ln P[\mathbf{r}|s] = T \sum_{a=1}^{N} r_a \ln\left(f_a(s)\right) + \ldots, \tag{3.30}$$

where the ellipsis represents terms that are independent or approximately independent of s, including, as discussed above, $\sum f_a(s)$. Setting the derivative to 0, we give the answer. □

Example 3.34. If the tuning curves are the Gaussians of Equation 3.27, this equation can be solved explicitly using the result $f_a'(s)/f_a(s) = (s_a - s)/\sigma_a^2$,

$$s_{\mathrm{ML}} = \frac{\sum r_a s_a/\sigma_a^2}{\sum r_a/\sigma_a^2}. \tag{3.31}$$

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

$$s_{\mathrm{ML}} = \frac{\sum r_a s_a}{\sum r_a}, \tag{3.32}$$

which is a simple estimation formula with an intuitive interpretation as the firing-rate weighted average of the preferred values of the encoding neurons.

Remark 3.20. Although the stimuli obtained by maximum likelihood estimation are the weighted average of the best responses. This result looks very good, but under the influence of noise, this method may reduce the accuracy. The MAP algorithm allows us to include prior knowledge $p[s]$ about the distribution of stimulus values in the decoding estimate. When using MAP, The objective function will have one more term $\log(p[s])$ and the maximum value can still be obtained by derivation. If the $p[s]$ is constant, the MAP and ML estimates are identical.

Inaddition, if many neurons are observed, or if a small number of neurons is observed over a long trial period, even a nonconstant stimulus distribution has little effect and $s_{\mathrm{MAP}} \approx s_{\mathrm{ML}}$.

Proposition 3.35. The MAP estimate is computed from the distribution $p[s|\mathbf{r}]$ determined by Bayes theorem. In terms of the logarithms of the probabilities, $\ln p[s|\mathbf{r}] = \ln p[\mathbf{r}|s] + \ln p[s] - \ln P[r]$. The last term in this expression is independent of s and can be absorbed into the ignored s-independent terms, so we can write, as in Equation 3.30,

$$\ln p[s|\mathbf{r}] = T\sum_{a=1}^{N} r_a \ln\left(f_a(s)\right) + \ln p[s] + \ldots. \tag{3.33}$$

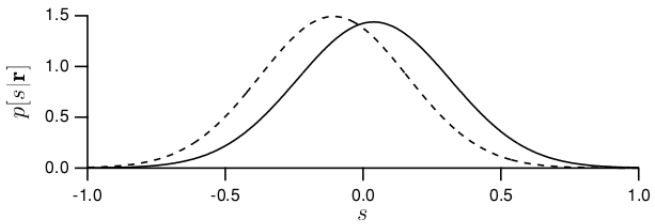
Maximizing this determines the MAP estimate,

$$T\sum_{a=1}^{N} \frac{r_a f_a'(s_{\rm MAP})}{f_a(s_{MAP})} + \frac{p'[s_{\rm MAP}]}{p[s_{\rm MAP}]} = 0. \tag{3.34}$$

Example 3.36. If the stimulus or prior distribution is itself Gaussian with mean $s_{\rm prior}$ and variance $\sigma_{\rm prior}$, and we use the Gaussian array of tuning curves, Equation 3.34 yields

$$s_{\rm MAP} = \frac{T\sum r_a s_a/\sigma_a^2 + s_{\rm prior}/\sigma_{\rm prior}^2}{T\sum r_a/\sigma_a^2 + 1/\sigma_{\rm prior}^2}. \tag{3.35}$$

Example 3.37. The figure compares the conditional stimulus probability densities $p[s|\mathbf{r}]$ for a constant stimulus distribution (solid curve) and for a Gaussian stimulus distribution with $s_{\rm prior} = -2$ and $\sigma_{\rm prior} = 1$, using the firing rates given by the filled circles in last figure. If the stimulus distribution is constant, $p[s|\mathbf{r}]$ peaks near the true stimulus value of 0. The effect of a noncon- stant stimulus distribution is to shift the curve toward the value -2, where the stimulus probability density has its maximum, and to decrease its width by a small amount. The estimate is shifted to the left because the prior distribution suggests that the stimulus is more likely to take negative values than positive ones, independent of the evoked response. The decreased width is due to the added information that the prior distribution provides.



Definition 3.38. The accuracy with which an estimate $s_{\rm est}$ describes a stimulus s can be characterized by two important quantities, its bias $b_{\rm est}(s)$ and its variance $\sigma^2_{\rm est}(s)$. The bias is the difference between the average of $s_{\rm est}$ across trials that use the stimulus s and the true value of the stimulus (i.e., s),

$$b_{\rm est}(s) = \langle s_{\rm est}\rangle - s. \tag{3.36}$$

Definition 3.39. An estimate is termed unbiased if $b_{\rm est}(s) = 0$ for all stimulus values.

Definition 3.40. The variance of the estimator, which quantifies how much the estimate varies about its mean value, is defined as

$$\sigma^2_{\rm est}(s) = \left\langle \left(s_{\rm est} - \langle s_{\rm est}\rangle\right)^2 \right\rangle. \tag{3.37}$$

Proposition 3.41. The bias and variance can be used to compute the trial-average squared estimation error, $\left\langle (s_{\rm est} - s)^2 \right\rangle$. This is a measure of the spread of the estimated values about the true value of the stimulus. Considering Definition 3.38, we can write the squared estimation error as

$$\left\langle (s_{\rm est} - s)^2 \right\rangle = \left\langle (s_{\rm est} - \langle s_{\rm est}\rangle + b_{\rm est}(s))^2 \right\rangle = \sigma^2_{\rm est}(s) + b^2_{\rm est}(s). \tag{3.38}$$

In other words, the average squared estimation error is the sum of the variance and the square of the bias. For an unbiased estimate, the average squared estimation error is equal to the variance of the estimator.

Remark 3.21. In general, minimizing the decoding error in Equation 3.38 involves a trade-off between minimizing the bias and minimizing the variance of the estimator.

3.3.3 Fisher Information

Remark 3.22. Decoding can be used to limit the accuracy with which a neural system encodes the value of a stimulus parameter because the encoding accuracy cannot exceed the accuracy of an optimal decoding method.

Definition 3.42. The *Fisher information* is a quantity that provides one such measure of encoding accuracy. Through a bound known as the *Cramér-Rao bound*, the Fisher information limits the accuracy with which any decoding scheme can extract an estimate of an encoded quantity.

Proposition 3.43. The Cramér-Rao lower bound for an estimator $s_{\rm est}$ is based on the Cauchy- Schwarz inequality, which states that for any two quantities A and B,

$$\langle A^2\rangle\langle B^2\rangle \geq \langle AB\rangle^2. \tag{3.39}$$

Proof. Note that

$$\left\langle \left(\langle B^2\rangle A - \langle AB\rangle B\right)^2 \right\rangle \geq 0 \tag{3.40}$$

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

$$\langle B^2\rangle^2\langle A^2\rangle - \langle AB\rangle^2\langle B^2\rangle \geq 0 \tag{3.41}$$

from which the inequality follows directly. □

Proposition 3.44. The *Cramér-Rao bound* limits the variance of any estimate $s_{\rm est}$ according to

$$\sigma^2_{\rm est}(s) \geq \frac{\left(1 + b'_{\rm est}(s)\right)^2}{I_{\rm F}(s)}, \tag{3.42}$$

where $b'_{\rm est}(s)$ is the derivative of $b_{\rm est}(s)$ and $I_{\rm F}(s)$ is the Fisher information.

Proof. Consider the inequality of Equation 3.39 with $A = \partial \ln p/\sigma^2_{\rm est}$ The Cauchy-Schwarz inequality then gives

$$\sigma^2_{\rm est}(s)I_F \geq \left\langle \frac{\partial \ln p[\mathbf{r}|s]}{\partial s}(s_{\rm est} - \langle s_{\rm est}\rangle)\right\rangle^2. \tag{3.43}$$

To evaluate the expression on the right side of the inequality 3.43), we differentiate the defining equation for the bias (Equation 3.36),

$$s + b_{\rm est}(s) = \langle s_{\rm est}\rangle = \int p[\mathbf{r}|s]s_{\rm est}d\mathbf{r}, \tag{3.44}$$

with respect to s to obtain

$$\begin{aligned} 1 + b'_{\rm est}(s) &= \int \frac{\partial p[\mathbf{r}|s]}{\partial s}s_{\rm est}d\mathbf{r} \\ &= \int p[\mathbf{r}|s]\frac{\partial \ln p[\mathbf{r}|s]\partial}{\partial s}s_{\rm est}d\mathbf{r} \\ &= \int p[\mathbf{r}|s]\frac{\partial \ln p[\mathbf{r}|s]\partial}{\partial s}(s_{\rm est} - \langle s_{\rm est}\rangle). \end{aligned} \tag{3.45}$$

The last equality follows from the identity

$$\int p[\mathbf{r}|s]\frac{\partial \ln p[\mathbf{r}|s]}{\partial s}\langle s_{\rm est}\rangle d\mathbf{r} = \langle s_{\rm est}\rangle \int \frac{\partial p[\mathbf{r}|s]}{\partial s}d\mathbf{r} = 0, \tag{3.46}$$

because $\int p[\mathbf{r}|s]d\mathbf{r} = 1$. The last line of equation 3.45 is just another way of writing the expression being squared on the right side of the inequality 3.43, so combining this result with the inequality gives

$$\sigma^2_{\rm est}(s)I_F \geq (1 + b'_{\rm est}(s))^2, \tag{3.47}$$

which, when rearranged, is the Cramér-Rao bound of Equation 3.42. □

Proposition 3.45. If we assume here that the firing rates take continuous values and that their distribution in response to a stimulus s is described by the conditional probability density $p[\mathbf{r}|s]$(assuming the latter is sufficiently smooth) by, $I_{\rm F}(s)$ can be written as

$$I_{\rm F}(s) = \left\langle -\frac{\partial^2 \ln p[\mathbf{r}|s]}{\partial s^2}\right\rangle = \int p[\mathbf{r}|s]\left(-\frac{\partial^2 \ln p[\mathbf{r}|s]}{\partial s^2}\right)d\mathbf{r}, \tag{3.48}$$

We can verify that the Fisher information can also be written as

$$\left\langle \left(\frac{\partial \ln p[\mathbf{r}|s]}{\partial s}\right)^2\right\rangle = \int p[\mathbf{r}|s]\left(\frac{\partial \ln p[\mathbf{r}|s]}{\partial s}\right)^2 d\mathbf{r}. \tag{3.49}$$

Remark 3.23. As Equation 3.48 shows, the Fisher information is a measure of the expected curvature of the log likelihood at the stimulus value s. Curvature is important because the likelihood is expected to be at a maximum near the true stimulus value s that caused the responses. Therefore, we can get two cases:

(i) If the likelihood is very curved, and thus the Fisher information is large, responses typical for the stimulus s are much less likely to occur for slightly different stimuli. Therefore, the typical response provides a strong indication of the value of the stimulus.

(ii) If the likelihood is fairly flat, and thus the Fisher information is small, responses common for s are likely to occur for slightly different stimuli as well. Thus, the response does not as clearly determine the stimulus value.

Remark 3.24. The Fisher information is purely local in the sense that it does not reflect the existence of stimulus values completely different from s that are likely to evoke the same responses as those evoked by s itself.

Remark 3.25. The Cramér-Rao bound sets a limit on the accuracy of any unbiased estimate of the stimulus. When $b_{\rm est}(s) = 0$, Equation 3.38 indicates that the average squared estimation error is equal to $\sigma^2_{\rm est}$ and, by Equation 3.42, this satisfies the bound $\sigma^2_{\rm est} \geq 1/I_{\rm F}(s)$. Provided that we restrict ourselves to unbiased decoding schemes, the Fisher information sets an absolute limit on decoding accuracy, and it thus provides a useful limit on encoding accuracy. In some cases, biased schemes may produce more accurate results than unbiased ones. For a biased estimator, the average squared estimation error and the variance of the estimate are not equal, and the estimation error can be either larger or smaller than $1/I_{\rm F}(s)$.

Remark 3.26. The limit on decoding accuracy set by the Fisher information can be attained by a decoding scheme we have studied, the maximum likelihood method. In the limit of large numbers of encoding neurons, and for most firing-rate distributions, the ML estimate is unbiased and saturates the Cramér-Rao bound.

Definition 3.46. Any unbiased estimator that saturates the Cramér-Rao lower bound is called efficient, i.e.

$$\sigma^2_{\rm est} = 1/I_{\rm F}(s). \tag{3.50}$$

Theorem 3.47. $I_{\rm F}(s)$ grows linearly with N, and the ML estimate obeys a central limit theorem, so that $N^{1/2}(s_{\rm ML} - s)$ is Gaussian distributed with a variance that is independent of N in the large N limit. In the limit $N \to \infty$, the ML estimate is asymptotically consistent, in the sense that $P[|s_{\rm ML} - s| > \varepsilon] \to 0$ for any $\varepsilon > 0$.

Example 3.48. The Fisher information for a population of neurons with uniformly arrayed tuning curves (the Gaussian array in example 3.30 for example) and Poisson statistics can be computed from the conditional firing-rate probability in Equation 3.30. Because the spike-count rate is described here by a probability rather than a probability density, we use the discrete analog of Equation 3.48,

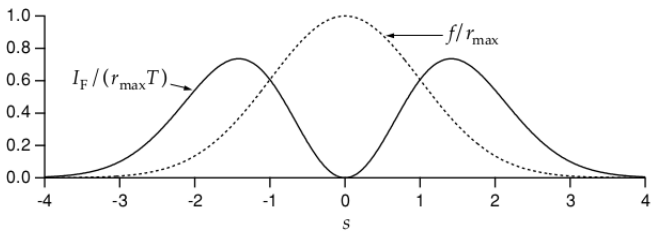
$$I_{\rm F}(s) = \left\langle -\frac{\partial^2 \ln p[\mathbf{r}|s]}{\partial s^2}\right\rangle = T\sum_{a=1}^{N}\langle r_a\rangle\left(\left(\frac{f'_a(s)}{f_a(s)}\right)^2 - \frac{f''_a(s)}{f_a(s)}\right). \tag{3.51}$$

If we assume that the array of tuning curves is symmetric, like the Gaussian array in example 3.30, the second term in the parentheses of the last expression sums to 0. We can also make the replacement $\langle r_a\rangle = f_a(s)$, producing the final result

$$I_{\rm F}(s) = T\sum_{a=1}^{N}\frac{f'_a(s)^2}{f_a(s)}. \tag{3.52}$$

In this expression, each neuron contributes an amount to the Fisher information proportional to the square of its tuning curve slope and inversely proportional to the average firing rate for the particular stimulus value being estimated.

Example 3.49. The Fisher information for a single neuron with a Gaussian tuning curve with $s = 0$ and $\sigma_a = 1$, 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 tuning curve is highest, and vanishes at $s = 0$, where the tuning curve peaks.



Remark 3.27. Individual neurons carry the most Fisher information in regions of their tuning curves where average firing rates are rapidly varying functions of the stimulus value, not where the firing rate is highest.

Remark 3.28. The Fisher information can be used to derive an interesting result on the optimal widths of response tuning curves. Consider a population of neurons with tuning curves of identical shapes, distributed evenly over a range of stimulus values as in Example 3.30. Equation 3.52 indicates that the Fisher information will be largest if the tuning curves of individual neurons are rapidly varying (making the square of their derivatives large), and if many neurons respond (making the sum over neurons large). For typical neuronal response tuning curves, these two requirements are in conflict. If the population of neurons has narrow tuning curves, individual neural responses are rapidly varying functions of the stimulus, but few neurons respond. Broad tuning curves allow many neurons to respond, but the individual responses are not as sensitive to the stimulus value.

Example 3.50. To determine whether narrow or broad tuning curves produce the more accurate encodings, we consider a dense distribution of Gaussian tuning curves, all with $\sigma_a = \sigma_r$. Using such curves in Equation 3.52, we find

$$I_F(s) = T\sum_{a=1}^{N} \frac{r_{\max}(s-s_a)^2}{\sigma_r^4} \exp\left(-\frac{1}{2}\left(\frac{s-s_a}{\sigma_r}\right)^2\right). \quad (3.53)$$

Proposition 3.51. This expression can be approximated by replacing the sum over neurons with an integral over their preferred stimulus values and multiplying by a density factor ρ_s. The factor ρ_s is the density with which the neurons cover the range of stimulus values, and it is equal to the number of neurons with preferred stimulus values lying within a unit range of s values. Replacing the sum over a with an integral over a continuous preferred stimulus parameter ξ (which replaces s_a), we find

$$\begin{aligned} I_F(s) &= \rho_s T \int_{-\infty}^{\infty} \frac{r_{\max}(s-\xi)^2}{\sigma_r^4} \exp\left(-\frac{1}{2}\left(\frac{s-\xi}{\sigma_r}\right)^2\right) \\ &= \frac{\sqrt{2\pi}\rho_s\sigma_r r_{\max}T}{\sigma_r^2}. \end{aligned} \quad (3.54)$$

The number of neurons that respond to a given stimulus value is roughly $\rho_s\sigma_r$, and the Fisher information is proportional to this number divided by the square of the tuning curve width seen from Equation 3.54.Combining these factors, the Fisher information is inversely proportional to σ_r, and the encoding accuracy increases with narrower tuning curves.

Proposition 3.52. Consider a stimulus with D parameters and suppose that the response tuning curves are products of identical Gaussians for each of these parameters. If the tuning curves cover the D-dimensional space of stimulus values with a uniform density ρ_s, the number of responding neurons for any stimulus value is proportional to $\rho_s\sigma_r^D$ and, using the same integral approximation as in equation (3.54), the Fisher information is

$$I_F = \frac{(2\pi)^{D/2}D\rho_s\sigma_r^D r_{\max}T}{\sigma_r^2} = (2\pi)^{D/2}D\rho_s\sigma_r^{D-2}r_{\max}T. \quad (3.55)$$

Remark 3.29. The trade-off 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_r smaller) increases the Fisher information for $D = 1$, decreases it for $D > 2$, and has no impact if $D = 2$.

3.3.4 Optimal Discrimination

Remark 3.30. 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.

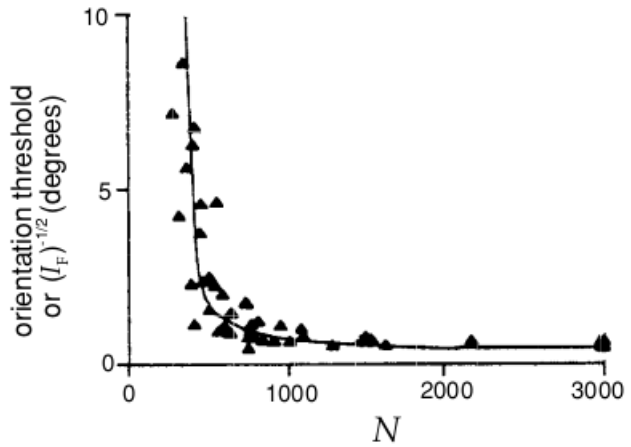
Proposition 3.53. Consider the case of discriminating between s and $s + \Delta s$ for small Δs. For large N, the average value of the difference between the ML estimates for the two stimulus values is equal to Δs (because the estimate is unbiased) and the variance of each estimate (for small Δs) is $1/I_F(s)$. Thus, the discriminability, defined in Equation 3.1, for the ML-based test is

$$d' = \Delta s\sqrt{I_F(s)}. \quad (3.56)$$

It can be known that the larger the Fisher information, the higher the discriminability.

Exercise 3.54. Proof that for small s, this discriminability is the same as that of the likelihood ratio test $Z(\mathbf{r})$ defined in Equation 3.54.

Example 3.55. The figure makes a 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.



3.4 Spike-Train Decoding

Remark 3.31. The decoding methods we have considered estimate or discriminate static stimulus values on the basis of spike-count firing rates. Spike-count 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 from the sequence of firing times t_i for $i = 1, 2, \cdots, n$ of the spikes that it evokes. Here we can solve this problem in two ways, one is to allow time lag, and the firing sequence after time t is also taken into account; the other is to introduce prediction, using the firing sequence before time t to predict the stimulus at time t.

Assumption 3.56. For simplicity, we restrict our discussion to the decoding of a single neuron. We assume, as we did in chapter 2, that the time average of the stimulus being estimated is 0.

Proposition 3.57. For the decoding of neurons, we may need to consider spikes for the following two time periods.

(i) The firing of an action potential at time t_i is only affected by the stimulus $s(t)$ prior to that time, $t < t_i$. That is, the evoked spikes tell us about the past behavior of the stimulus,and if stimulus have some form of temporal correlation, that past behavior provides information about the current stimulus value. Of course, if there is no correlation between the stimulus s, then there is no need to consider the firing sequence before time t.

(ii) The firing sequence after time t is the neuron's encoding of the stimulus at time t and the previous stimulus, so we must consider.

Proposition 3.58. 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 could have been fired in response to the stimulus $s(t - \tau_0)$ being estimated (those for which $t - \tau_0 < t_i < t$), and spikes (those for which $t_i < t - \tau_0$) which can contribute to its estimation on the basis of stimulus correlations.

Remark 3.32. 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.

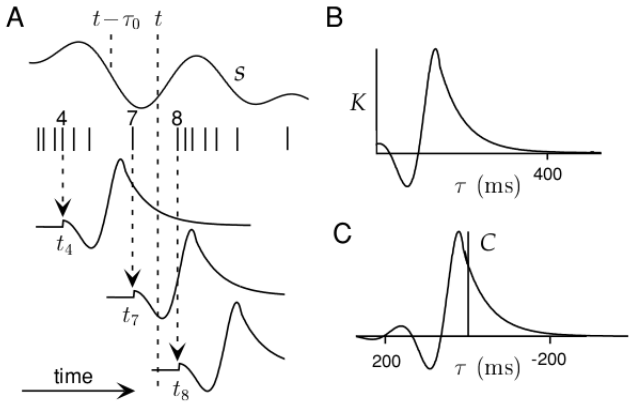
Proposition 3.59. The stimulus estimate is constructed as a linear sum over all spikes. A spike occurring at time t_i contributes a kernel $K(t - t_i)$, and the total estimate is obtained by summing over all spikes,

$$s_{\text{est}(t-\tau_0)} = \sum_{i=1}^{n} K(t - t_i) - \langle r \rangle \int_{-\infty}^{\infty} K d\tau. \tag{3.57}$$

The last term, with $\langle r \rangle = \langle n \rangle / T$ the average firing rate over the trial, is included to impose the condition that the time average s_{est} is 0, in agreement with the time-average condition on s.

Definition 3.60. A kernel be imposed by requiring $K(t - t_0) = 0$ for $(t - t_i) \leq 0$ is termed causal.

Example 3.61. The sum in Equation 3.57 includes all spikes, according to Proposition 3.59 so only those spikes occurring prior to the time t (spikes $1 - 7$ in figure) should be included. The estimate is obtained by summing the values of the kernels where they cross the dashed line labeled t, for spikes up to and including spike 7. The causal kernel is 0 for negative values of its argument, so spikes for $i \geq 8$ do not contribute to the estimate at this time. Figure A shows how spikes contribute to a stimulus estimate,using the kernel shown in figure B. For figure C, we will explain in the following example.



Assumption 3.62. We ignore the causality constraint for now and construct an acausal kernel, but we will return to issues of causality later in the discussion.

Proposition 3.63. Equation 3.57 can be written in a compact way by using the neural response function $\rho(t)\sum\delta(t-t_i)$ introduced in chapter 1,

$$s_{\text{est}(t-\tau_0)} = \int_{\infty}^{\infty} (\rho(t-\tau) - \langle r\rangle)K(\tau)d\tau. \tag{3.58}$$

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 2.

Proposition 3.64. We choose K so that the squared difference between the estimated stimulus and the actual stimulus, averaged over both time and trials,

$$\frac{1}{T}\int_0^T \left\langle \left(\int_{-\infty}^{\infty} (\rho(t-\tau) - \langle r\rangle)K(\tau) - s(t-\tau_0)\right)^2 \right\rangle d\tau, \tag{3.59}$$

is minimized.

Proof. Using the calculus of variations,the result is that optimal kernal K obeys the equation

$$\int_{\infty}^{\infty} Q_{\rho\rho}(\tau-\tau')K(\tau')d\tau' = Q_{\text{rs}}(\tau-\tau_0), \tag{3.60}$$

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

$$Q_{\rho\rho} = \frac{1}{T}\int_0^T \langle(\rho(t-\tau) - \langle r\rangle)(\rho(t-\tau') - \langle r\rangle)\rangle\, dt, \tag{3.61}$$

as defined in chapter 1. Q_{rs} is the correlation of the firing rate and the stimulus, which is related to the spike-triggered average C, both introduced in chapter 1,

$$Q_{\text{rs}}(\tau-\tau_0) = \langle r\rangle\, C(\tau_0-\tau) = \frac{1}{T}\left\langle \sum_{i=1}^{n} s(t_i+\tau-\tau_0)\right\rangle, \tag{3.62}$$

which completes the proof. □

Example 3.65. If the spike train is uncorrelated, which tends to happen at low rates,

$$Q_{\rho\rho} = \langle r\rangle\,\delta(\tau), \tag{3.63}$$

and we find from Equation 3.60 that

$$\begin{aligned} K(\tau) &= \frac{1}{\langle r\rangle}Q_{\text{rs}}(\tau-\tau_0) = C(\tau_0-\tau) \\ &= \frac{1}{\langle n\rangle}\left\langle \sum_{i=1}^{n} s(t_i+\tau-\tau_0)\right\rangle. \end{aligned} \tag{3.64}$$

This is the average value of the stimulus at time $\tau-\tau_0$ relative to the appearance of a spike. The figure C in Example 3.61 is the spike-triggered average corresponding to the kernel shown in figure B, assuming no spike-train correlations.

Remark 3.33. Because $\tau-\tau_0$ can be either positive or negative, stimulus estimation involves both forward and backward correlation and the average values of the stimulus both before and after a spike. Decoding in this way follows a simple rule: every time a spike appears, we replace it with the average stimulus surrounding a spike, shifted by an amount τ_0.

Remark 3.34. Correlations between a spike and subsequent stimuli can arise, in a causal system, only from correlations between the stimulus and itself. If these are absent, as for white noise, $K(\tau)$ will be 0 for $\tau > \tau_0$. For causal decoding, we must also have $K(\tau) = 0$ for $\tau < 0$. Thus, if $\tau_0 = 0$ and the stimulus is uncorrelated, $K(\tau) = 0$ for all values of τ.

Proposition 3.66. When the spike-train autocorrelation function is not a δ function, an acausal solution for K can be expressed as an inverse Fourier transform,

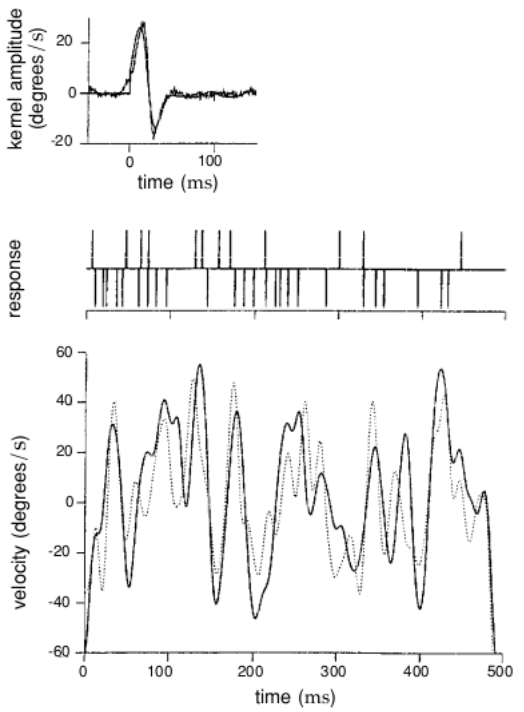
$$K(\tau) = \frac{1}{2\pi}\int \tilde{K}(\omega)\exp(-i\omega\tau), \tag{3.65}$$

where

$$\tilde{K}(\omega) = \frac{\tilde{Q}_{\text{rs}}(\omega)\exp(-i\omega\tau_0)}{\tilde{Q}_{\rho\rho}(\omega)}. \tag{3.66}$$

Here $\tilde{Q}_{\text{rs}}$ and $\tilde{Q}_{\rho\rho}$ are the Fourier transforms of Q_{rs} and $Q_{\rho\rho}$.

Example 3.67. The figure shows an example of spike-train decoding for the H1 neuron of the fly discussed in chapter 2. Flies have two H1 neurons, one on each side of the body, that respond to motion in opposite directions. As is often the case, half-wave rectification prevents a single neuron from encoding both directions of motion. The top panel gives two reconstruction kernels, one acausal and one causal. The two rows of spikes in the middle panel shows typical responses of the H1 neuron to the stimuli (upper trace) $s(t)$ and its negative, $-s(t)$(bottom trace). This procedure provides a reasonable approximation of recording both H1 neurons, and produces a neuron/anti-neuron pair of recordings. The stimulus is then decoded by summing the kernel $K(t-t_i)$ for all spike times t_i of the recorded H1 neuron and summing $-K(t-t_i)$ for all spike times t_j of its anti-neuron partner. The dashed line in the lower panel shows the actual stimulus, and the solid line is the estimated stimulus from the optimal linear reconstruction using the acausal filter.



Chapter 4

Information Theory

4.1 Entropy and Mutual Information

Remark 4.1. 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.

Definition 4.1. *Information theory* is a general framework for quantifying the ability of a coding scheme or a communication channel (such as the optic nerve) to convey information.

Assumption 4.2. In information theory, it is assumed that the code involves a number of symbols, and that the coding and transmission processes are stochastic and noisy.

Definition 4.3. *Entropy* is a measure of the theoretical capacity of a code to convey information.

Definition 4.4. *Mutual information* is a measure of how much of that capacity is actually used when the code is employed to describe a particular set of data.

Remark 4.2. Communication channels, if they are noisy, have only limited capacities to convey information. The techniques of information theory are used to evaluate these limits and find coding schemes that saturate them.

Notation 17. In neuroscience applications, the symbols we consider are neuronal responses, and the data sets they describe are stimulus characteristics. We discuss cases in which the symbols consist of responses described by spike-count firing rates r as the simplified descriptions of the response of a neuron that reduce the number of possible "symbols" (i.e., responses) that need to be considered. In the following, we just consider neuroscience applications.

Remark 4.3. Because a reduced description of a spike train can carry no more information than the full spike train itself, this approach provides a lower bound on the actual information carried by the spike train.

4.1.1 Entropy

Fact 4.5. Entropy is a quantity that, roughly speaking, measures how "interesting" or "surprising" a set of responses is.

Example 4.6. The most widely used measure of entropy, due to Shannon, expresses the "surprise" associated with seeing a response rate r as a function of the probability of getting that response, $h(P[r])$, and quantifies the entropy as the average of $h(P[r])$ over all possible responses. The function $h(P[r])$, which acts as a measure of surprise, is chosen to satisfy a number of conditions:

(1) $h(P[r])$ should be a decreasing function of $P[r]$ because low probability responses are more surprising than high probability responses.

(2) The surprise measure for a response that consists of two independent spike counts should be the sum of the measures for each spike count separately. Suppose we record rates r_1 and r_2 from two neurons that respond independently of each other, the additivity condition requires that

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

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

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

where the minus sign makes h a decreasing function of its argument as required.

Remark 4.4. 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.

Definition 4.7. Equation 4.2 quantifies the surprise or unpredictability associated with a particular response. *Shannon's entropy* is this measure averaged entropy over all responses,

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

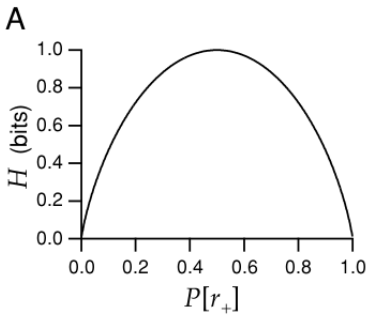
where r is spike-count firing rates (i.e., the number of spikes divided by the trial duration).

Notation 18. In the following, the entropy means Shannon's entropy.

Example 4.8. The neuron responds in only two possible ways, either with rate r_+ or r_-. In this case, there are only two nonzero terms in Equation 4.3, and, using the fact that $P[r_-] = 1 - P[r_+]$, the entropy is

$$H = -(1 - P[r_+]) \log_2(1 - P[r_+]) - P[r_+] \log_2(P[r_+]).$$

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



4.1.2 Mutual Information

Remark 4.5. Entropy is a measure of response variability, but it does not tell us anything about the source of that variability. A neuron can provide information about a stimulus only if its response variability is correlated with changes in that stimulus, rather than being purely random or correlated with other unrelated factors. Mutual information is an entropy-based measure related to this idea.

Theorem 4.9. The entropy of the responses to a given stimulus s is

$$H_s = -\sum_r P[r|s] \log_2 P[r|s]. \tag{4.4}$$

Proof. The entropy of the responses evoked by repeated presentations of a given stimulus s is computed using the conditional probability $P[r|s]$, the probability of a response at rate r given that stimulus s. This proof is completed by Definition 4.7. □

Definition 4.10. The *noise entropy* 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. It can be obtained by averaging the quantity 4.4 over all the stimuli,

$$\begin{aligned} H_{\text{noise}} &= \sum_s P[s] H_s \\ &= -\sum_{s,r} P[s] P[r|s] \log_2 P[r|s]. \end{aligned} \tag{4.5}$$

Definition 4.11. The *mutual information* is the difference between the total response entropy and the noise entropy, which from equations 4.3 and 4.5 gives

$$\begin{aligned} I_{\text{m}} &= H - H_{\text{noise}} \\ &= -\sum_r P[r] \log_2 P[r] + \sum_{s,r} P[s] P[r|s] \log_2 P[r|s]. \end{aligned} \tag{4.6}$$

Proposition 4.12. The mutul information defined in Equation 4.6 can be written in two forms as follows,

$$I_{\text{m}} = \sum_{s,r} P[s] P[r|s] \log_2 \left(\frac{P[r|s]}{P[r]} \right), \tag{4.7}$$

$$I_{\text{m}} = \sum_{s,r} P[r,s] \log_2 \left(\frac{P[r,s]}{P[r]P[s]} \right). \tag{4.8}$$

Proof. The first equation can be derived by using

$$P[r] = \sum_s P[s] P[r|s], \tag{4.9}$$

and writing the difference of the two logarithms in Equation 4.6 as the logarithm of the ratio of their arguments. Recall from chapter 3 that

$$P[r,s] = P[s]P[r|s] = P[r]P[s|r], \tag{4.10}$$

where $P[r,s]$ is the joint probability of stimulus s appearing and response r being evoked. This equation can be used to derive the second form of the above equations for the mutual information. □

Remark 4.6. Equation 4.8 reveals that the mutual information is symmetric with respect to interchange of s and r, which means that the mutual information that a set of responses conveys about a set of stimuli is identical to the mutual information that the set of stimuli conveys about the responses.

Theorem 4.13. The mutul information also satisfies

$$I_{\text{m}} = -\sum_s P[s] \log_2 P[s] + \sum_{s,r} P[r] P[s|r] \log_2 P[s|r], \tag{4.11}$$

which is the same as Equation 4.6, except that the roles of the stimulus and the response have been interchanged.

Proof. Applying the second equality of Equation 4.10 in Equation 4.8 completes this proof. □

Remark 4.7. Equation 4.11 shows how response variability limits the ability of a spike train to carry information. The second term on the right side, which is negative, is the average uncertainty about the identity of the stimulus given the response, and reduces the total stimulus entropy represented by the first term.

Example 4.14. Suppose that the responses of the neuron are completely unaffected by the identity of the stimulus. In this case, $P[r|s] = P[r]$, and from Equation 4.7 it follows immediately that $I_{\rm m} = 0$.

Example 4.15. Suppose that each stimulus s produces a unique and distinct response r_s. Then, $P[r_s] = P[s]$ and $P[r|s]$ is 1 if $r = r_s$ and 0 otherwise. This causes the sum over r in Equation 4.7 to collapse to just one term, and the mutual information becomes

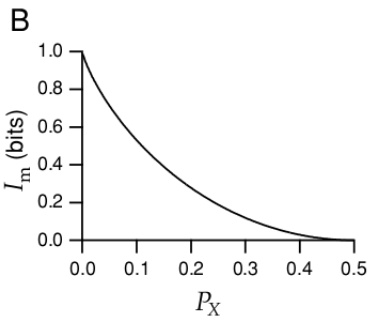
$$I_{\rm m} = \sum_s P[s] \log_2 \left(\frac{1}{P[r_s]} \right) = -\sum_s P[s] \log_2 P[s], \quad (4.12)$$

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

Example 4.16. Imagine that there are only two possible stimulus values, which we label + and −, and that the neuron responds with just two rates, r_+ and r_-. We associate the response r_+ with the + stimulus, and the response r_- with the − stimulus, but the encoding is not perfect. The probability of an incorrect response is P_X, meaning that for the correct responses $P[r_+|+] = P[r_-|-] = 1 - P_X$, and for the incorrect responses $P[r_+|-] = P[r_-|+] = P_X$. We assume that the two stimuli are presented with equal probability so that $P[r_+] = P[r_-] = 1/2$, which makes the full response entropy 1 bit. The noise entropy is $-(1 - P_X) \log_2(1 - P_X) - P_X \log_2 P_X$. Thus, the mutual information is

$$I_{\rm m} = 1 + (1 - P_X) \log_2(1 - P_X) + P_X \log_2 P_X, \quad (4.13)$$

which is plotted in the following figure. When the encoding is error-free ($P_X = 0$), the mutual information is 1 bit, which is equal to both the full response entropy and the stimulus entropy. When the encoding is random ($P_X = 1/2$), the mutual information goes to 0.



Remark 4.8. It is instructive to consider this example from the perspective of decoding. We can think of the neuron as being a communication channel that reports noisily on the stimulus. From this perspective, we want to know the probability that a + was presented, given that the response r_+ was recorded. By Bayes theorem, this is $P[+|r_+] = P[r_+|+]P[+]/P[r_+] = 1 - P_X$. Before the response is recorded, the expectation was that + and − were equally likely. If the response r_+ is recorded, this expectation changes to $1 - P_X$. The mutual information measures the corresponding reduction in uncertainty or, equivalently, the tightening of the posterior distribution due to the response.

Remark 4.9. The mutual information is related to a measure used in statistics called the Kullback-Leibler (KL) divergence.

Definition 4.17. The *Kullback-Leibler(KL) divergence* between one probability distribution $P[r]$ and another distribution $Q[r]$ is

$$D_{\rm KL}(P, Q) = \sum_r P[r] \log_2 \left(\frac{P[r]}{Q[r]} \right). \quad (4.14)$$

Theorem 4.18. The KL divergence has a property normally associated with a distance measure, $D_{\rm KL}(P, Q) \geq 0$ with equality if and only if $P = Q$.

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

$$\begin{aligned} -D_{\rm KL}(P, Q) &= \sum_r P[r] \log_2 \left(\frac{Q[r]}{P[r]} \right) \\ &\leq \log_2 \left(\sum_r P[r] \frac{Q[r]}{P[r]} \right) = 0. \end{aligned} \quad (4.15)$$

The last equality holds because $Q[r]$ is a probability distribution and thus satisfies $\sum_r Q[r] = 1$. Equation 4.15 implies that $D_{\rm KL}(P, Q) \geq 0$, with equality holding if and only if $P[r] = Q[r]$. □

Corollary 4.19. A similar result holds for the Kullback-Leibler divergence between two probability densities,

$$D_{\rm KL}(p, q) = \int p[r] \log_2 \frac{p[r]}{q[r]} dr \geq 0. \quad (4.16)$$

Theorem 4.20. The mutual information is the KL divergence between the distributions $P[r, s]$ and $P[r]P[s]$.

Proof. This is directly from Definition 4.17 and Equation 4.8. □

Remark 4.10. If the stimulus and the response were independent of one another, $P[r, s]$ would be equal to $P[r]P[s]$. Thus, the mutual information is the KL divergence between the actual probability distribution $P[r, s]$ and the value it would take if the stimulus and response were independent.

Corollary 4.21. 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.

Proof. The first conclusion is directly from theorems 4.20 and 4.18. The second is from equations 4.6 and 4.11. □

4.1.3 Entropy and Mutual Information for Continuous Variables

Remark 4.11. Up to now we have characterized neural responses using discrete spike count rates. As in chapter 3, it is often convenient to treat these rates instead as continuous variables.

Fact 4.22. If we could measure the value of a continuously defined firing rate with unlimited accuracy, it would be possible to convey an infinite amount of information using the endless sequence of decimal digits of this single variable. Of course, practical considerations always limit the accuracy with which a firing rate can be measured or conveyed.

Remark 4.12. To define the entropy associated with a continuous measure of a neural response, we must include some limit on the measurement accuracy.

Proposition 4.23. For a continuously defined firing rate, the entropy with measurement accuracy Δr satisfies

$$H = -\sum p[r]\Delta r \log_2 p[r] - \log_2 \Delta r. \tag{4.17}$$

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

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

where the last equality follows from the fact that the sum of the response probabilities is 1. □

Remark 4.13. 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.

Theorem 4.24. In the limit $\Delta r \to 0$ with the sum replaced by an integral, we can write

$$\lim_{\Delta r \to 0} \{H + \log_2 \Delta r\} = -\int p[r] \log_2 p[r] dr, \tag{4.18}$$

where Δr is best thought of as a limit on the resolution with which the firing rate can be measured.

Remark 4.14. If two entropies computed with the same resolution are subtracted, the troublesome term involving Δr cancels, and we can proceed without knowing its precise value.

Definition 4.25. The integral on the right side of Equation 4.18 is sometimes called the *differential entropy*.

Proposition 4.26. The noise entropy, for a continuous variable like the firing rate, can be written in a manner similar to the response entropy Equation 4.18, except that the conditional probability density $p[r|s]$ is used:

$$\lim_{\Delta r \to 0} \{H_{\text{noise}} + \log_2 \Delta r\} = -\iint p[s]p[r|s] \log_2 p[r|s] dr ds. \tag{4.19}$$

Proposition 4.27. The mutual information is the difference between the expressions in equations 4.18 and 4.19,

$$I_{\text{m}} = \iint p[s]p[r|s] \log_2 \left(\frac{p[r|s]}{p[r]} \right) dr ds. \tag{4.20}$$

Proof. Here the factor of $\log_2 \Delta r$ cancels because both entropies are evaluated at the same resolution. □

Remark 4.15. There are some differences and relationshipsbetween Fisher information and mutual information:

(i) We described the Fisher information as a local measure of how tightly the responses determine the stimulus. The Fisher information is local because it depends on the expected curvature of the likelihood $P[\mathbf{r}|s]$ (typically for the responses of many cells) evaluated at the true stimulus value.

(ii) The mutual information is a global measure in the sense that it depends on the average overall uncertainty in the decoding distribution $p[s|\mathbf{r}]$, including values of s both close to and far from the true stimulu s. If the decoding distribution $p[s|\mathbf{r}]$ has a single peak about the true stimulus, the Fisher information and the mutual information are closely related. ??????????????????????????????????????

4.2 Information and Entropy Maximization

Remark 4.16. 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.

Remark 4.17. Because the mutual information is the full response entropy minus the noise entropy, maximizing the information involves a compromise. We must make the response entropy as large as possible without allowing the noise entropy to get too big. If the noise entropy is small, maximizing the response entropy, subject to an appropriate constraint, maximizes the mutual information to a good approximation. We therefore begin our discussion by studying how response entropy can be maximized. Later in the discussion, we will consider the effects of noise entropy.

Remark 4.18. Constraints play a crucial role in this analysis. We have already seen that the theoretical information-carrying capacity associated with a continuous firing rate is limited only by the resolution with which the firing rate can be defined.??????????????????????????????

4.2.1 Entropy Maximization for a Single Neuron

Assumption 4.28. During the maximization process, the resolution 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.

Remark 4.19. To maximize the response entropy, we must find a probability density $p[r]$ that makes the integral in equation 4.18 as large as possible while satisfying whatever constraints we impose. One constraint that always applies in entropy maximization is that the integral of the probability density must be 1.

Proposition 4.29. Suppose that the neuron in question has a maximum firing rate of $r_{\max}$. Then,

$$p[r] = \frac{1}{r_{\max}} \tag{4.21}$$

solves

$$\max_{p[r]} \int_0^{r_{\max}} p[r] \log_2 p[r], \ s.t. \ \int_0^{r_{\max}} p[r] dr = 1. \tag{4.22}$$

The entropy for this probability density, for finite firing rate resolution Δr, is

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

Proof. Equation 4.21 is from Lagrange multipliers and the same way to compute the functional derivative as Proposition 2.8. Thus, $H = \log_2 r_{\max} - \log_2 \Delta r = \log_2 \left(\frac{r_{\max}}{\Delta r} \right)$. □

Remark 4.20. Equation 4.21 is independent of r and is the basis of a signal-processing technique called *histogram equalization*. Applied to neural responses, this is a procedure for tailoring the neuronal selectivity so that $p[r] = 1/r_{\max}$ in response to a set of stimuli over which the entropy is to be maximized.

Proposition 4.30. Suppose a neuron responds to a stimulus characterized by the parameter s by firing at a rate $r = f(s)$. If the response probability density takes its optimal value, $p[r] = 1/r_{\max}$, the corresponding turning curve in the case of a monotonically increasing response is

$$f(s) = r_{\max} \int_{s_{\min}}^{s} p[s'] ds', \tag{4.24}$$

where $s_{\min}$ is the minimum value of s, which is assumed to generate no response.

Proof. For small Δs, the probability that the continuous stimulus variable falls in the range between s and $s + \Delta s$ is given in terms of the stimulus probability density by $p[s]\Delta s$. This produces a response that falls in the range between $f(s + \Delta s)$ and $f(s)$. When $p[r] = 1/r_{\max}$, 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

$$\frac{|f(s + \Delta s) - f(s)|}{r_{\max}} = p[s]\Delta s. \tag{4.25}$$

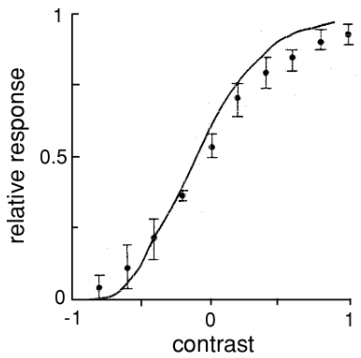
Because of the monotonically increasing response so that $f(s+\Delta s) > f(s)$ for positive Δs. Then,in the limit $\Delta s \to 0$, the equalization condition becomes

$$\frac{df}{ds} = r_{\max} p[s], \tag{4.26}$$

which has the solution Equation 4.24. □

Remark 4.21. In the case of Proposition 4.30, entropy maximization requires that the average firing rate of the responding neuron be proportional to the integral of the probability density of the stimulus.

Example 4.31. The large monopolar cell (LMC) in the visual system of the fly responds to contrast, and the following figure shows that contrast response of the fly LMC (data points) compared to the integral of the natural contrast probability distribution (solid curve). The relative response is the amplitude of the membrane potential fluctuation produced by the onset of a light or dark image with a given level of contrast divided by the maximum response. Contrast is defined relative to the background level of illumination.



The response as a function of contrast is very close to the integrated probability density, suggesting that the LMC is using a maximum entropy encoding.

Remark 4.22. These responses in Example 4.31 are measured as membrane potential fluctuation amplitudes, not as firing rates, but the analysis presented can be applied without modification.

Remark 4.23. 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 the variance of the firing rate might be constrained.

Exercise 4.32. Show that the entropy-maximizing probability density is an exponential, if the average firing rate is constrained to a fixed value.

Exercise 4.33. Show that the probability density that maximizes the entropy subject to constraints on the firing rate and its variance is a Gaussian.

4.2.2 Populations of Neurons

Remark 4.24. When a population of neurons encodes a stimulus, optimizing their individual response properties will not necessarily lead to an optimized population response. Optimizing individual responses could result in a highly redundant population representation in which different neurons encode the same information.

Fact 4.34. Entropy maximization for a population requires that the neurons convey independent pieces of information (i.e., they must have different response selectivities).

Notation 19. Let the vector $\mathbf{r}$ with components r_a for $a = 1, 2, \cdots, N$ denote the firing rates for a population of N neurons, measured with resolution Δr.

Theorem 4.35. 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

$$H = -\int p[\mathbf{r}] \log_2 p[r] d\mathbf{r} - N \log_2 \Delta r. \tag{4.27}$$

Proposition 4.36. Along with the full population entropy of Equation 4.27, we can also consider the entropy associated with individual neurons within the population. If $p[r_a] = \int \prod_{b \neq a} p[\mathbf{r}] dr_b$ is the probability density for response r_a from neuron a, its entropy is

$$\begin{aligned} H_a &= -\int p[r_a] \log_2 p[r_a] dr_a - \log_2 \Delta r \\ &= -\int p[\mathbf{r}] \log_2 p[r_a] d\mathbf{r} - \log_2 \Delta r. \end{aligned} \tag{4.28}$$

Proposition 4.37. The true population entropy can never be greater than the sum of these individual neuron entropies over the entire population,

$$H \leq \sum_a H_a. \tag{4.29}$$

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

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

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

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

that is, if the responses of the neurons are statistically independent. Thus, the full response entropy is never greater than the sum of the entropies of the individual neurons in the population, and it reaches the limiting value when Equation 4.30 is satisfied. □

Definition 4.38. A code that satisfies this condition is called a *factorial code* because the probability factorizes into a product of single neuron probabilities.

Remark 4.25. When the population-response probability density factorizes, this implies that the individual neurons respond independently. The entropy difference in Equation 4.37 has been suggested as a measure of redundancy.

Proposition 4.39. Combining the result in proposition 4.37 with the results of the previous section, we con- clude that the maximum population-response entropy can be achieved by satisfying two conditions:

1. The individual neurons must respond in- dependently, which means that $p[\mathbf{r}] = \prod_a p[r_a]$ must factorize.
2. If the same constraint is imposed on every neuron, the second condition implies that every neuron must have the same response probability density. In other words, $p[r_a]$ must be the same for all a values, a property called probability equalization.

Remark 4.26. We proceed by considering factorization and probability equalization as general principles of entropy maximization, without imposing explicit constraints.

Remark 4.27. 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?????????.

Theorem 4.40. A more modest goal is to require that the lowest-order moments of the population-response probability distribution match those of a fully factorized and equalized distribution. If the individual response probability distributions are equal, the average firing rates and firing rate variances will be the same for all neurons.

$$\langle r_a \rangle = \langle r \rangle \quad \text{and} \quad \left\langle (r_a - \langle r \rangle)^2 \right\rangle = \sigma_r^2. \tag{4.31}$$

for all a. Furthermore, the covariance matrix for a factorized and probability-equalized population distribution is proportional to the identity matrix,

$$Q_{ab} = \int p[\mathbf{r}] \left(r_a - \langle r \rangle\right) \left(r_b - \langle r \rangle\right) d\mathbf{r} = \sigma_r^2 \delta_{ab}. \tag{4.32}$$

Remark 4.28. Finding response distributions that satisfy only the decorrelation and vari- ance equalization condition of equation 4.31 is usually tractable. In the fol- lowing examples, we restrict ourselves to this easier task. This maximizes the entropy only if the statistics of the responses are Gaussian, but it is a reasonable procedure even in a non-Gaussian case, because it typically reduces the redundancy in the population code and spreads the load of information transmission equally among the neurons.

4.2.3 Application to Retinal Ganglion Cell Receptive Fields

Assumption 4.41. Entropy and information maximization have been used to explain proper- ties of visual receptive fields in the retina, LGN, and primary visual cor- tex. The basic assumption is that these receptive fields serve to maximize the amount of information that the associated neural responses convey about natural visual scenes in the presence of noise.

Remark 4.29. Information theoret- ical 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 re- dundancy. Maximizing the information conveyed requires removing this redundancy from the neural responses.

Remark 4.30. It should be kept in mind that the information maximization approach sets limited goals and requires strong assumptions about the nature of the constraints relevant to the nervous system. In addition, the approach analyzes only the representational properties of neural responses and ignores the computational goals of the visual system, such as object recognition or target tracking. Never- theless, the principle of information maximization is quite successful at accounting for properties of receptive fields early in the visual pathway.

Notation 20. In chapter 2, a visual image was defined by a contrast function $s(x, y, t)$ with a trial-averaged value of 0. For the calculations we present here, it is more convenient to express the x and y coordinates for locations on the viewing screen in terms of a single vector $\vec{x} = (x, y)$, or sometimes $\vec{y} = (x, y)$.

Theorem 4.42. The linear estimate of the response of a visual neuron discussed in chapter 2 can be written as

$$L(t) = \int_0^\infty \int D(\vec{x}, \tau) s(\vec{x}, t - \tau) d\vec{x} d\tau. \tag{4.33}$$

Proposition 4.43. If the space-time receptive field $D(\vec{x}, \tau)$ is separable, $D(\vec{x}, \tau) = D_s(\vec{x}) D_t(\tau)$, and we can rewrite $L(t)$ as the product of integrals involving temporal and spatial filters. To keep the notation simple, we assume that the stimulus can also be separated, so that $s(\vec{x}, \tau) = s_s(\vec{x}) s_t(t)$. Then,

$$L(t) = L_s L_t(t), \tag{4.34}$$

where

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

and

$$L_t(t) = \int_0^\infty D_t(\tau) s_t(t - \tau) d\tau. \tag{4.36}$$

Remark 4.31. In the following, we analyze the spatial and temporal components, D_s and D_t, separately by considering the information-carrying capacity of L_s and L_t.

Assumption 4.44. 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 which is most effective are spatially (and temporally) stationary or translation-invariant. This means that all locations and directions in space (and all times), at least within the patch we con- sider, 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.

Notation 21. Note that we are labeling the neurons by the locations $\vec{a}$ of the centers of their receptive fields rather than by an integer index such as i. This is a convenient labeling scheme that allows sums over neurons to be replaced by sums over parameters describing their receptive fields.The vectors $\vec{a}$ for the different neurons take on discrete values corresponding to the different neurons in the population.

Proposition 4.45. Based on the above assumptions, 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

$$L_s(\vec{a}) = \int D_s(\vec{x}) - \vec{a} s_s(\vec{x}) d\vec{x}. \tag{4.37}$$

Remark 4.32. If many neurons are being considered, these discrete vectors may fill the range of receptive field locations quite densely. In this case, it is reasonable to approximate the large but discrete set of $\vec{a}$ values with a vector $\vec{a}$ that is allowed to vary continuously. In other words, as an approximation, we proceed as if there were a neuron corresponding to every continuous value of $\vec{a}$. This allows us to treat as a function of $\vec{a}$ and to replace sums over neurons with integrals over $\vec{a}$. In the case we are considering, the receptive fields of retinal ganglion cells cover the retina densely, with many receptive fields overlapping each point on the retina, so the replacement of discrete sums over neurons with continuous integrals over $\vec{a}$ is quite accurate.

4.2.4 The Whitening Filter

The relevant correlation is the average, over all stimuli, of the product of the linear responses of two cells, with receptive fields centered at $\vec{a}$ and $\vec{b}$,

$$\begin{aligned} Q_{\mathrm{LL}}(\vec{a}, \vec{b}) &= \left\langle L_s(\vec{a}) L_s(\vec{b}) \right\rangle \\ &= \int D_s(\vec{x}) - \vec{a}) D_s(\vec{y}) - \vec{b}) \left\langle s_s(\vec{x}) s_s(\vec{y}) \right\rangle d\vec{x} d\vec{y}. \end{aligned} \tag{4.38}$$

Here the average, denoted by angle brackets, is not over trials but over the set of natural scenes for which we believe the receptive field is optimized.

Proposition 4.46. By analogy with Equation 4.32, decorrelation and variance equalization of the different retinal ganglion cells, when $\vec{a}$ and $\vec{a}$ are taken to be continuous variables, require that we set this correlation function proportional to a δ function,

$$Q_{\mathrm{LL}}(\vec{a}, \vec{b}) = \sigma_L^2 \delta(\vec{a} - \vec{b}). \tag{4.39}$$

which is the continuous variable analog of making a discrete correlation ma- trix proportional to the identity matrix (Equation 4.32).

Theorem 4.47. Our assumption of homogeneity 4.44 implies that $\langle s_s(\vec{x}) s_s(\vec{y}) \rangle$ is only a function of the vector difference $\vec{x} - \vec{y}$ (actually, if all directions are equivalent, it is only a function of the magnitude $\vec{x} - \vec{y}$), and we write it as

$$Q_{ss}(\vec{x} - \vec{y}) = \langle s_s(\vec{x}) s_s(\vec{y}) \rangle . \tag{4.40}$$

Theorem 4.48. The optimal receptive field filter (entropy maximization) for receptive field filter is

$$\left|\tilde{D}_s(\vec{\kappa})\right| = \frac{\sigma_L}{\sqrt{\tilde{Q}_{ss}(\vec{\kappa})}}. \tag{4.41}$$

Proof. To determine the form of the receptive field filter that is optimal, we must solve equation 4.39 for D_s. This is done by expressing D_s and Q_{ss} in terms of their Fourier transforms $\tilde{D}_s$ an $\tilde{Q}_{ss}$,

$$\begin{aligned} D_s(\vec{x}-\vec{a}) &= \frac{1}{4\pi^2}\int \exp\left(-i\vec{\kappa}\cdot(\vec{x}-\vec{a})\right)\tilde{D}_s(\vec{\kappa})d\vec{\kappa} \\ QQ_{ss}(\vec{x}-\vec{y}) &= \frac{1}{4\pi^2}\int \exp\left(-i\vec{\kappa}\cdot(\vec{x}-\vec{y})\right)\tilde{Q}_{ss}(\vec{\kappa})d\vec{\kappa}. \end{aligned} \tag{4.42}$$

where $\tilde{Q}_{ss}$ is real and nonnegative, is also called the stimulus power spectrum. In terms of these Fourier transforms, Equation 4.39 becomes

$$\left|\tilde{D}_s(\vec{\kappa})\right|^2 \tilde{Q}_{ss}(\vec{\kappa}) = \sigma_L^2. \tag{4.43}$$

□

Remark 4.33. The linear kernel described by equation 4.42 exactly compensates for whatever dependence the Fourier transform of the stimulus correlation function has on the spatial frequency $\vec{\kappa}$, making the product $\tilde{Q}_{ss}(\vec{\kappa})\left|\tilde{D}_s(\vec{\kappa})\right|^2$ independent of $\vec{\kappa}$. This product is the power spectrum of L.

Definition 4.49. The output of the optimal filter has a power spectrum that is independent of spatial fre- quency, and therefore has the same characteristics as white noise. There- fore, the kernel in Equation 4.41 is called a whitening filter.

Remark 4.34. Different spatial frequencies act independently in a linear system,??????????????????????

Remark 4.35. The calculation we have performed determines only the amplitude $\left|\tilde{D}_s(\vec{\kappa})\right|$ and not $\tilde{D}_s(\vec{\kappa})$ itself. Thus, decorrelation and variance equalization do not uniquely specify the form of the linear kernel. We study some consequences of the freedom to choose different linear kernels satisfying Equation 4.41 later in the chapter.

Example 4.50. The spatial correlation function for natural scenes has been measured, with the result that $\tilde{Q}_{ss}(\vec{\kappa})$ is proportional to $1/\left|\vec{\kappa}\right|^2$ over the range it has been evaluated. The behavior near $\vec{\kappa}=0$ is not well established, but the divergence of $1/\left|\vec{\kappa}\right|^2$ near $\vec{\kappa}=0$ can be removed by setting $\tilde{Q}_{ss}(\vec{\kappa})$ proportional to $1/(\left|\vec{\kappa}\right|^2+\kappa_0^2)$ where κ_0 is a constant. The stimuli of interest in the calculation of retinal ganglion receptive fields are natural images as they appear on the retina, not in the photographs from which the natural scenes statistics are measured. An additional factor must be included in $\tilde{Q}_{ss}(\vec{\kappa})$ to account for filtering introduced by the optics of the eye (the optical modulation transfer function). A simple model of the optical modulation transfer function results in an exponential correction to the stimulus correlation function,

$$\tilde{Q}_{ss}(\vec{\kappa}) \propto \frac{\exp(-\alpha\left|\vec{\kappa}\right|)}{\left|\vec{\kappa}\right|^2+\kappa_0^2}, \tag{4.44}$$

with α a parameter. Substituting this into Equation 4.41 gives the rather eculiar result that the amplitude $\tilde{D}_s(\vec{\kappa})$, being proportional to the inverse of the square root of $\tilde{Q}_{ss}$, is predicted to grow exponentially for large $|\vec{\kappa}|$.

Theorem 4.51. *Whitening filters* maximize entropy by equalizing the distribution of response power over the entire spatial frequency range.

Remark 4.36. High spatial frequency components of images are relatively rare in natural scenes and, even if they occur, are greatly attenuated by the eye. The whitening filter compensates for this by boosting the responses to high spatial frequencies. Although this is the result of the entropy maximization calculation, it is not a good strategy to use in an unrestricted way for visual processing.

Remark 4.37. Real inputs to retinal ganglion cells involve a mixture of true signal and noise coming from biophysical sources in the retina. At high spatial frequencies, for which the true signal is weak, inputs to retinal ganglion cells are likely to be dominated by noise, especially in low-light conditions. Boosting the amplitude of this noise-dominated input and transmitting it to the brain is not an efficient visual encoding strategy. The problem of excessive boosting of responses at high spatial frequency arises in the entropy maximization calculation because no distinction has been made between the entropy coming from true signals and that coming from noise.

4.2.5 Filtering Input Noise

Remark 4.38. To correct the problem caused by noise, we should maximize the information transmitted by the retinal ganglion cells about natural scenes, rather than maximize the entropy. A full information-maximization calculation of the receptive field properties of retinal ganglion cells can be performed, but this requires introducing a number of assumptions about the constraints that are relevant, and it is not entirely obvious what these constraints should be. Instead, we will follow an approximate procedure that prefilters the input to eliminate as much noise as possible, and then uses the results of this section to maximize the entropy of a linear filter acting on the prefiltered input signal

Assumption 4.52. Suppose that the visual stimulus on the retina is the sum of the true stimulus $s_s\vec{x}$ that should be conveyed to the brain and a noise term $\eta(\vec{x})$ that reflects image distortion, photoreceptor noise, and other signals that are not worth conveying beyond the retina.

Theorem 4.53. To deal with such a mixed input signal, we express the Fourier transform of the linear kernel $\tilde{D}_s(\vec{\kappa})$ as a product of two terms: a noise filter, $\tilde{D}_\eta(\vec{\kappa})$, that eliminates as much of the noise as possible; and a whitening

filter, $\tilde{D}_w(\vec{\kappa})$, that satisfies Equation (4.41). The Fourier transform of the complete filter is then

$$\tilde{D}_s(\vec{\kappa}) = \tilde{D}_w(\vec{\kappa})\tilde{D}_\eta(\vec{\kappa}). \tag{4.45}$$

Assumption 4.54. we assume that the signal and noise terms are uncorrelated, so that $\langle s_s(\vec{x})\eta(\vec{y})\rangle = 0$.

Theorem 4.55. The optimal noise filter is real and given, in terms of the Fourier transforms of Q_{ss} and $Q_{\eta\eta}$, by

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

Proof. To determine the form of the noise filter, we demand that when it is applied to the total input $s_s(\vec{x}) + \eta(\vec{x})$, the result is as close to the signal part of the input, $s_s(\vec{x})$, as possible. The relevant cross- correlation for this problem is

$$\langle (s_s(\vec{x}) + \eta(\vec{x}))s_s(\vec{y})\rangle = Q_{ss}(\vec{x} - \vec{y}), \tag{4.47}$$

and the autocorrelation is

$$\langle (s_s(\vec{x}) + \eta(\vec{x}))(s_s(\vec{y}) + \eta(\vec{y}))\rangle = Q_{ss}(\vec{x} - \vec{y}) + Q_{\eta\eta}(\vec{x} - \vec{y}), \tag{4.48}$$

where Q_{ss} and $Q_{\eta\eta}$ are, respectively, the stimulus and noise autocorrelations functions. The problem is to minimize the average squared difference between the filtered noisy signal and the true signal. The general solution is that the Fourier transform of the optimal filter is the Fourier transform of the cross-correlation between the quantity being filtered and the quantity being approximated divided by the Fourier transform of the autocorrelation of the quantity being filtered. □

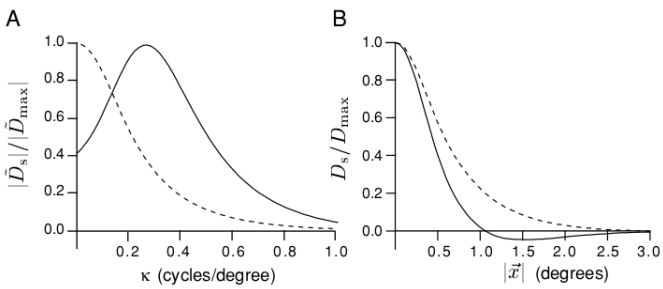
Theorem 4.56. the noise filter is designed so that its output matches the signal as closely as possible, we make the approximation of using the same whiten- ing filter as before (Equation 4.41). Combining the two, we find that

$$\left|\tilde{D}_s(\vec{\kappa})\right| \propto \frac{\sigma_L\sqrt{\tilde{Q}_{ss}(\vec{\kappa})}}{\tilde{Q}_{ss}(\vec{\kappa}) + \tilde{Q}_{\eta\eta}(\vec{\kappa})}. \tag{4.49}$$

Example 4.57. Linear kernels resulting from Equation 4.49, using Equation 4.44 for the stimulus correlation function, are plotted in figure A and B. For these figures, we have assumed that the input noise is white so that $\tilde{Q}_{\eta\eta}(\vec{\kappa})$ is independent of $\vec{\kappa}$. The two figures descirbes are receptive field properties predicted by entropy maximization and noise suppression of responses to natural images.

Both the amplitude of the Fourier transform of the kernel (figure A) and the actual spatial kernel $\tilde{D}_s(\vec{\kappa})$ (figure B) are plotted under conditions of low and high noise. The linear kernels in figure B have been constructed by assuming that $\tilde{D}_s(\vec{\kappa})$ satisfies Equation 4.49 and is real, which minimizes the spatial extent of the resulting receptive field. The resulting function $D_s(\vec{x})$ is radially symmetric, so it depends only on the distance $|\vec{x}|$ from the center of the receptive field to the point $\vec{x}$, and this radial dependence is plotted in figure B. Under low noise conditions (solid lines in figure), the linear kernel has a bandpass character and the predicted receptive field has a center-surround structure, which matches the retinal ganglion receptive fields shown in chapter 2.

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



4.2.6 Temporal Processing in the LGN

Remark 4.39. Natural images tend to change relatively slowly over time. This means that there is substantial redundancy in the succession of natural images, suggesting an opportunity for efficient temporal filtering to complement efficient spatial filtering.

Remark 4.40. 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.

Theorem 4.58. Recall that the predicted linear temporal response is given by $L_t(t)$, as expressed in Equation 4.36.The analog of Equation 4.39 for temporal decorrelation and variance equalization is

$$\langle L_t(t)L_t(t')\rangle = \sigma_L^2\delta(t - t'). \tag{4.50}$$

which is mathematically identical to equation 4.39 except that the role of the spatial variables $\vec{a}$ and $\vec{b}$ has been replaced by the temporal variables t and t'.

Theorem 4.59. 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,

$$D_t(\tau) = \frac{1}{2\pi}\int \exp(-i\omega\tau)\tilde{D}_t(\omega)d\omega, \tag{4.51}$$

and $\tilde{D}_t(\omega)$ is given by an equation similar to Equation 4.49,

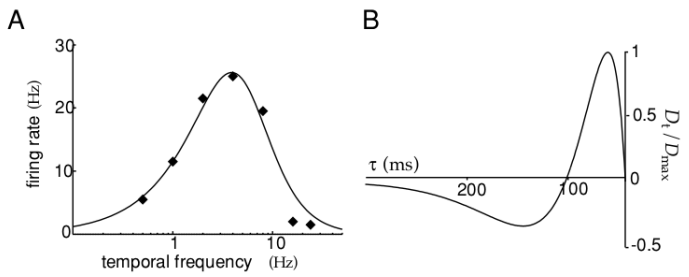
$$\left|\tilde{D}_t(\omega)\right| \propto \frac{\sigma_L \sqrt{\tilde{Q}_{ss}(\omega)}}{\tilde{Q}_{ss}(\omega) + \tilde{Q}_{\eta\eta}(\omega)}. \tag{4.52}$$

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

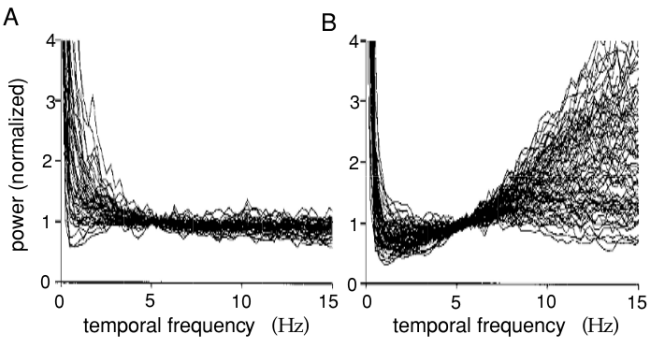
Example 4.60. Dong and Atick (1995) analyzed temporal receptive fields in the LGN in this way, under the assumption that a substantial fraction of the temporal redundancy of visual stimuli is removed in the LGN rather than in the retina. They determined that the temporal power spectrum of natural scenes has the form

$$\tilde{Q}_{ss}(\omega) \propto \frac{1}{\omega^2 + \omega_0^2}, \tag{4.53}$$

where ω is a constant. The resulting filter, in both the temporal frequency and the time domains, is plotted in figure. Figure A shows the pre- dicted and actual frequency responses of an LGN cell. Because the optimiza- tion procedure determines only the amplitude of the Fourier transform of the linear kernel, $D_t(\tau)$ is not uniquely specified. To determine the tem- poral kernel, we require it to be causal ($D_t(\tau) = 0 \quad \text{for} \quad \tau < 0$) and impose a technical condition known as minimum phase?????????????, which assures that the out- put changes as rapidly as possible when the stimulus varies. Figure B shows the resulting form of the temporal filter. Figure B shows the resulting form of the temporal filter. The space-time receptive fields shown in chapter 2 tend to change sign as a function of τ. The tem- poral filter in figure B has exactly this property.



Example 4.61. 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 A shows the power spectra of spike trains of cat LGN cells in response to natural scenes (the movie Casablanca), and figure B shows power spectra in response to white-noise stimuli. The power spectra of the responses to natural scenes are quite flat above about $\omega = 3$ Hz. In response to white noise, on the other hand, they rise with ω. This is exactly what we would expect if LGN cells are acting as temporal whitening filters????????. In the case of natural stimuli, the whitening filter evenly distributes the output power over a broad fre- quency range. Responses to white-noise stimuli increase at high frequen- cies due to the boosting of inputs at these frequencies by the whitening filter???????.



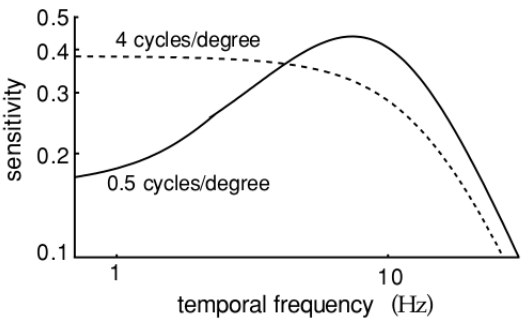
4.2.7 Cortical Coding

Remark 4.41. Computational concerns beyond mere linear information transfer are likely to be relevant at the level of cortical processing of visual images. For one thing, the primary visual cortex has many more neurons than the LGN, yet they can collectively convey no more information about the visual world than they receive. As we saw in chapter 2, neurons in primary visual cortex are selective for quantities, such as spatial frequency and orientation, that are of particular importance in relation to object recognition but not for information transfer.

Remark 4.42. The methods described in the previous section can be used to understand restricted aspects of receptive fields of neurons in primary visual cortex, namely, the way that their multiple selectivities are collectively assigned. For example, cells that respond best at high spatial frequencies tend to respond more to low temporal frequency components of images, and vice versa.

Example 4.62. The stimulus power spectrum written as a function of both spatial and temporal frequency has been estimated as $\tilde{Q}_{ss}(\vec{\kappa}, \omega) \propto 1/(|\vec{\kappa}|^2 + \alpha^2\omega^2)$,where $\alpha = 0.4$ cycle seconds/degree.This correlation function decreases both for high spatial and high temporal frequencies. The figure shows how temporal selectivity for a combined noise and whitening filter, constructed using this stimulus power spectrum, changes for different preferred spa- tial frequencies. The basic idea is that components with fairly low stimulus power are boosted by the whitening filter, while those with very low stim- ulus power get suppressed by the noise filter. As shown by Li (1996), if a cell is selective for high spatial frequencies, the input signal rapidly falls below the noise (treated as white) as the temporal frequency of the input is increased. As a result, the noise filter of Equation 4.46 causes the temporal response to be largest at 0 temporal frequency (dashed curve of figure). If instead the cell is selective for low spatial frequencies, the signal dominates the noise up to higher temporal frequencies, and the whitening filter causes the response to increase as a function of temporal frequency up to a maximum value where the noise filter begins to suppress the response (solid curve in figure)????????.

Receptive fields with preference for high spatial frequency thus act as low-pass temporal filters, and receptive fields with selectivity for low spatial frequency act as bandpass temporal filters.



Example 4.63. Similar conclusions can be drawn concerning other joint selectivities. For example, color-selective (chrominance) cells tend to be selective for low temporal frequencies, because their input signal-to-noise ratio is lower than that for broadband (luminance) cells. There is also an interesting predicted relationship between ocular dominance and spatial frequency tuning due to the nature of the correlations between the two eyes. Opti- mal receptive fields with low spatial frequency tuning (for which the in- put signal-to-noise ratio is high) have enhanced sensitivity to differences between inputs coming from the two eyes. Receptive fields tuned to inter- mediate and high spatial frequencies suppress ocular differences?????.

4.3 Entropy and Information for Spike Trains

Remark 4.43. Computing the entropy or information content of a neuronal response characterized by spike times is much more difficult than computing these quantities for responses described by firing rates. Nevertheless, these computations are important, because firing rates are incomplete descriptions that can lead to serious underestimates of the entropy and information.

Fact 4.64. Spike-train entropy calculations are typically based on the study of long-duration recordings consisting of many action potentials. The longer the total length of a spike train, the more information it contains.

Remark 4.44. By Fact 4.64, the entropy and mutual information of spike trains are reported as entropy or information rates.

Definition 4.65. The *entropy rate* and *information rate* are defined as the total entropy and information divided by the duration of the spike train, respectively. Alternatively, entropy and mutual information can be divided by the total number of action potentials and reported as bits per spike rather than bits per second.

Notation 22. We write the entropy rate as $\dot{H}$ rather than H.

Fact 4.66. The temporal pattern of a group of action potentials can be specified by listing either the individual spike times or the sequence of intervals between successive spikes.

Remark 4.45. The entropy and mutual information calculations we present are based on a spike-time description, but as an initial example we consider an approximate computation of entropy using interspike intervals.

4.3.1 Based on Interspike Intervals

Remark 4.46. The interspike interval is a continuous variable.

Notation 23. The probability of an interspike interval falling in the range between τ and $\tau+\Delta\tau$ is given in terms of the interspike interval probability density by $p[\tau]\Delta\tau$, where $\Delta\tau$ is the resolution.

Proposition 4.67. If the different interspike intervals are statistically independent and identically distributed, the entropy associated with the interspike intervals in a spike train of average rate $\langle r\rangle$ and of duration T is

$$H = -\langle r\rangle T \int_0^\infty p[\tau]\log_2(p[\tau]\Delta\tau)d\tau,$$

where $\langle r\rangle T$ is the number of intervals. In this case, the entropy rate is

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

Proof. These are directly from definitions of the entropy and entropy rate. □

Example 4.68. If a spike train is described by a homogeneous Poisson process with rate $\langle r\rangle$, we have

$$p[\tau] = \langle r\rangle\, e^{-\langle r\rangle\tau}$$

and the interspikes are statistically independent (Chapter 1). Thus,

$$\dot{H} = \frac{\langle r\rangle}{\ln 2}(1-\ln\langle r\rangle\,\Delta\tau). \tag{4.54}$$

In fact,

$$\begin{aligned}\dot{H} &= -\langle r\rangle \int_0^\infty \langle r\rangle\, e^{-\langle r\rangle\tau}\log_2(\langle r\rangle\, e^{-\langle r\rangle\tau}\Delta\tau)d\tau\\ &= -\langle r\rangle \int_0^\infty e^{-\tau}\log_2(\langle r\rangle\, e^{-\tau}\Delta\tau)d\tau\\ &= -\langle r\rangle\left(\log_2(\langle r\rangle\,\Delta\tau) + \int_0^\infty \frac{-e^{-\tau}}{\ln 2}d\tau\right)\\ &= \frac{\langle r\rangle}{\ln 2}(1-\ln\langle r\rangle\,\Delta\tau),\end{aligned}$$

where the second step follows from the variable substitution $\tau = \langle r\rangle\,\tau$ and the third step from the integration by parts.

Definition 4.69. Equation 4.54 is called the *Poisson entropy rate*.

Theorem 4.70. In general, the entropy rate $\dot{H}$ for a spike train with interspike interval distribution $p[\tau]$ and average rate $\langle r\rangle$ satisfies

$$\dot{H} \le -\langle r\rangle \int_0^\infty p[\tau]\log_2(p[\tau]\Delta\tau)d\tau. \tag{4.55}$$

Proof. Correlations between different interspike intervals reduce the total entropy, so the result obtained by assuming independent intervals provides an upper bound on the true entropy of a spike train. □

4.3.2 General Computations

Formula 4.71. To make entropy calculations practical, a long spike train is broken into statistically independent subunits, and the total entropy is written as the sum of the entropies for the individual subunits.

Example 4.72. In the case of Proposition 4.67, the subunit was the interspike interval.

Remark 4.47. If interspike intervals are not independent, and we wish to compute a result and not merely a bound, we must work with larger subunit descriptions.

Notation 24. The variable T_s is used below to denote the duration of the spike sequence being considered, while T, which is much larger than T_s, is the duration of the entire spike train.

Formula 4.73. Denote these basic subunits by spike sequences of duration T_s. A spike sequence can be obtained as follows.

(i) Divide time T_s into discrete bins of size Δt, which is small enough so that not more than one spike appears in a bin.

(ii) Label each bin by a 0 (no spike) or a 1 (spike), depending on whether or not a spike occurred within it.

(iii) Represent a spike sequence defined over a block of duration T_s by a string of $T_s/\Delta t$ zeros and ones.

We denote such a sequence by $B(t)$, where B is a $T_s/\Delta t$ bit binary number, and t specifies the time of the first bin in the sequence being considered. Both T_s and t are integer multiples of the bin size Δt.

Notation 25. The probability of a sequence B occurring at any time during the entire response is denoted by $P[B]$.

Remark 4.48. $P[B]$ can be obtained by counting the number of times the sequence B occurs anywhere within the spike trains being analyzed (*including overlapping cases*).

Proposition 4.74. The spike-train entropy rate implied by the distribution that is characterized by $P[B]$ is

$$\dot{H} = -\frac{1}{T_s}\sum_B P[B]\log_2 P[B], \tag{4.56}$$

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

Proposition 4.75. If the spike sequences in nonoverlapping intervals of duration T_s are independent and identically distributed, the full spike-train entropy rate is also given by Equation 4.56.

Proof. By the independence,

$$\begin{aligned}\dot{H} &= \frac{-T/T_s\sum_B P[B]\log_2 P[B]}{T}\\ &= -\frac{1}{T_s}\sum_B P[B]\log_2 P[B],\end{aligned}$$

which completes the proof. □

Theorem 4.76. For small T_s such that the spike sequences are not independent, Equation 4.56 provides an upper bound on the true entropy rate, that is,

$$\dot{H} \le -\frac{1}{T_s}\sum_B P[B]\log_2 P[B]. \tag{4.57}$$

Proof. Any correlations between successive intervals (if $B(t+T_s)$ is correlated with $B(t)$, for example) reduce the total spike-train entropy, causing Equation 4.56 to overestimate the true entropy rate. □

Remark 4.49. If T_s is too small, $B(t+T_s)$ and $B(t)$ are likely to be correlated, and the overestimate may be severe. As T_s increases, we expect the correlations to get smaller, and Equation 4.56 should provide a more accurate value.

Remark 4.50. For any finite data set, T_s cannot be increased past a certain point, because there will not be enough spike sequences of duration T_s in the data set to determine their probabilities. Thus, in practice, T_s must be increased until the point where the extraction of probabilities becomes problematic, and some form of extrapolation to $T_s \to \infty$ must be made.

Assumption 4.77. Statistical mechanics arguments suggest that the difference between the entropy rate for finite T_s and the true entropy rate for $T_s \to \infty$ should be proportional to $1/T_s$ for large T_s.

Proposition 4.78. The true entropy rate can be estimated by linearly extrapolating a plot of the entropy rate versus $1/T_s$ to the point $1/T_s = 0$.

Proof. This is directly from Assumption 4.77. □

Remark 4.51. To compute the mutual information rate for a spike train, we must subtract the full noise entropy rate from the full spike-train entropy rate.

Notation 26. $P[B(t)]$ is the probability of finding a given sequence B at time t within a set of spike trains obtained on trials using the same stimulus. In contrast, $P[B]$, used in the spike-train entropy rate calculation, is the probability of finding the sequence B at any time within these trains.

Lemma 4.79. If the same stimulus is used in repeated trials, the noise entropy rate at time t satisfies

$$\dot{H}_t = -\frac{1}{T_s}\sum_B P[B(t)]\log_2 P[B(t)].$$

Proof. The noise entropy rate is determined from the probabilities of finding various sequences B, given that they were evoked by the same stimulus. If the same stimulus is used in repeated trials, sequences $B(t)$ that begin at time t in every trial are generated by the same stimulus. Therefore, the conditional probability of the response, given the stimulus, is in this case the distribution $P[B(t)]$ for response sequences beginning at time t. This is obtained by determining the fraction of trials on which $B(t)$ was evoked. □

Remark 4.52. Determining $P[B(t)]$ for a sufficient number of spike sequences may take a large number of trials using the same stimulus.

Proposition 4.80. The full noise entropy rate can be computed by averaging the noise entropy rate at time t over all t values, that is,

$$\dot{H}_{noise} = -\frac{\Delta t}{T}\sum_t \left(\frac{1}{T_s}\sum_B P[B(t)]\log_2 P[B(t)]\right), \quad (4.58)$$

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

Proof. In this case, the average over t plays the role of the average over stimuli in Equation 4.5. Then, Lemma 4.79 completes the proof. □

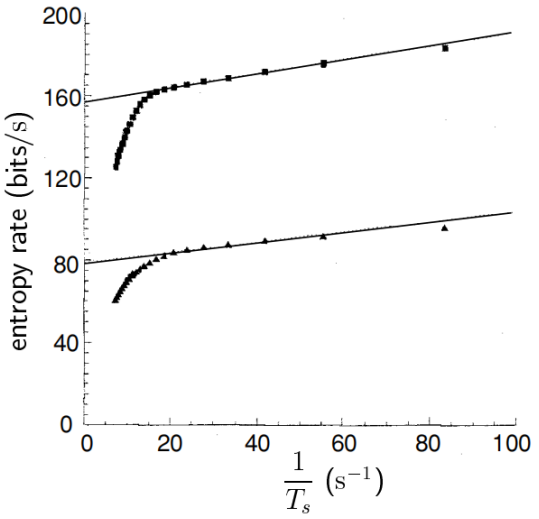
Theorem 4.81. If Equation 4.58 is based on finite-length spike sequences, it provides an upper bound on the noise entropy rate, that is,

$$\dot{H}_{noise} \leq -\frac{\Delta t}{T}\sum_t \left(\frac{1}{T_s}\sum_B P[B(t)]\log_2 P[B(t)]\right). \quad (4.59)$$

Proposition 4.82. The true noise entropy rate is estimated by performing a linear extrapolation in $1/T_s$ to $1/T_s = 0$.

Proof. As was done for the spike-train entropy rate. □

Example 4.83. Entropy and noise entropy rates for the H1 visual neuron in the fly responding to a randomly moving visual image are shown in the following picture. (i) The filled circles in the upper trace show the full spike-train entropy rate computed for different values of $1/T_s$. The straight line is a linear extrapolation to $1/T_s = 0$, which corresponds to $T_s \to \infty$. (ii) The lower trace shows the spike train noise entropy rate for different values of $1/T_s$. The straight line is again an extrapolation to $1/T_s = 0$.



Both entropy rates increase as functions of $1/T_s$, and the true spike-train and noise entropy rates are overestimated at large values of $1/T_s$. At $1/T_s \approx 20/s$, there is a sudden shift in the dependence. This occurs when there is insufficient data to compute the spike sequence probabilities. By linearly extrapolating the linear part of the series of computed points spike trains had an approximate entropy rate of 157 bits/s and an appeoximate noise entropy rate of 79 bits/s when the resolution was $\Delta t = 3$ ms. The information rate is obtained by taking the difference between the extrapolated values for the spiketrain and noise entropy rates. The result is an information rate of 157 - 79 = 78 bits/s or 1.8 bits/spike.

Remark 4.53. Both the spike-train and noise entropy rates depend on Δt. The leading dependence, coming from the $\log_2 \Delta t$ term discussed previously, cancels in the computation of the information rate, but the information can still depend on Δt through nondivergent terms. This reflects the fact that more information can be extracted from accurately measured spike times than from poorly measured spike times. Thus, we expect the information rate to increase with decreasing Δt, at least over some range of Δt values. At some critical value of Δt that matches the natural degree of noise jitter in the spike timings, we expect the information rate to stop increasing. This value of Δt is interesting because it tells us about the degree of spike timing accuracy in neural encoding.

Remark 4.54. The information conveyed by spike trains can be used to compare responses to different stimuli and thereby reveal stimulus-specific aspects of neural encoding.