Neural coding: linear models

Sebastian Seung

9.29 Lecture 1: February 4, 2003

1   What is computational neuroscience?

The term "computational neuroscience" has two different definitions:

1. using a computer to study the brain

2. studying the brain as a computer

In the first, the field is defined by a technique. In the second, it is defined by an idea. Let's discuss these two definitions in more depth.

Why use a computer to study the brain? The most compelling reason is the torrential flow of data generated by neurophysiology experiments. Today it is common to simultaneously record the signals generated by tens of neurons in an awake behaving animal. Once the measurement is done, the neuroscientist must analyze the data to figure out what it means, and computers are necessary for this task. Computers are also used to simulate neural systems. This is important when the models are complex, so that their behaviors are not obvious from mere verbal reasoning.

On to the second definition. What does it mean to say that the brain is a computer? To grasp this idea we must think beyond our desktop computers with their glowing screens. The abacus is a computer, and so is a slide rule. What do these examples have in common? They are all dynamical systems, but they are of a special class. What's special is that the state of a computer *represents* something else. The states of transistors in your computer's display memory represent the words and pictures that are displayed on its screen. The locations of the beads on a abacus represent the money passing through a shopkeeper's hands. And the activities of neurons in our brains represent the things that we sense and think about. In short,

$$\text{computation} = \text{coding} + \text{dynamics}$$

The two terms on the right hand side of this equation are the two great questions for computational neuroscience. How are computational variables are encoded in neural activity? How do the dynamical behaviors of neural networks emerge from the properties of neurons?

The first half of this course will address the problem of encoding, or representation. The second half of the course will address the issue of brain dynamics, but only incompletely. The biophysics of single neurons will be discussed, but the collective behaviors of networks are left for my other class 9.641 Neural Networks.

2 Neural coding

As an introduction to the problem of neural coding, let me show you a video of a neurophysiology experiment. This video comes from the laboratory of David Hubel, who won the Nobel prize with his colleague Torsten Wiesel for their discoveries in the mammalian visual system.

In the video, you will see a visual stimulus, a flashed or moving bar of light projected onto a screen. This is the stimulus that is being presented to the cat. You will also hear the activity of a neuron recorded from the cat's brain. I should also describe what you will *not* see and hear. A cat has been anesthetized and placed in front of the screen, with its eyelids held open. The tip of a tungsten wire has been placed inside the skull, and lodged next to a neuron in a visual area of the brain. Although the cat is not conscious, neurons in this area are still responsive to visual stimuli. The tungsten wire is connected to an amplifier, so that the weak electrical signals from the neuron can be recorded. The amplified signal is also used to drive a loudspeaker, and that is the sound that you will hear.

As played on the loudspeaker, the response of the neuron consists of brief clicking sounds. These clicks are due to spikes in the waveform of the electrical signal from the neuron. The more technical term for spike is action potential. Almost without exception, such spikes are characteristic of neural activity in the vertebrate brain.

As you can see and hear, the frequency of spiking is dependent on the properties of the stimulus. The neuron is activated only when the bar is placed at a particular location in the visual field. Furthermore, it is most strongly activated when the bar is presented at a particular orientation. Arriving at such a verbal model of neural coding is more difficult than it may seem from the video. David Hubel has recounted his feelings of frustration during his initial studies of the visual cortex. For a long time, he used spots of light as visual stimuli, because that had worked well in his previous studies of other visual areas of the brain. But spots of light evoked only feeble responses from cortical neurons. The spots of light were produced by a kind of slide projector. One day Hubel was wrapping up yet another unsuccessful experiment. As he pulled the slide out of the projector, he heard an eruption of spikes from the neuron. It was that observation that led to the discovery that cortical neurons were most sensitive to oriented stimuli like edges or bars.

The study of neural coding is not restricted to sensory processing. One can also investigate the neural coding of motor variables. In this video, you will see the movements of a goldfish eye, and hear the activity of a neuron involved in control of these movements. The oculomotor behavior consists of periods of static fixation, punctuated by rapid saccadic movements. The rate of action potential firing during the fixation periods is correlated with the horizontal position of the eye.

Finally, some neuroscientists study the encoding of computational variables that can't be classified as either sensory nor motor. This video shows a recording of a neuron in a rat as it moves about a circular arena. Neurons like this are sensitive to the direction of the rat's head relative to the arena, and are thought to be important for the rat's ability to navigate.

Verbal models are the first step towards understanding neural coding. But computational neuroscientists do not stop there. They strive for a deeper understanding by

constructing mathematically precise, quantitative models of neural coding. In the next few lectures, you will learn how to construct such models. But first you have to become familiar with the format of data from neurophysiological experiments.

3   Outline of the first part of the class

1. convolution and correlation

2. Wiener-Hopf equations

3. visual receptive fields

4. fourier analysis and the auditory system

5. probabilistic models of spike trains

4   Discretely sampled data

For your first homework assignment, you will be given data from an experiment on the weakly electric fish *Eigenmannia*. The fish has a special organ that generates an oscillating electric field with a frequency of several hundred Hz. It also has an electrosensory organ, with which it is able to sense its electric field and the fields of other fish. The electric field is used for electrolocation and communication.

In the experiment, the fish was stimulated with an artificial electric field, and the activity of a neuron in the electrosensory organ was recorded. The artificial electric field was an amplitude-modulated sine wave, much like the natural electric field of the fish. The stimulus vector s_i in the dataset contains the modulation signal sampled every 0.5 ms.

The extracellular voltage waveform was an analog signal. In such a recording, action potentials are submillivolt events. Some sort of pattern detector is applied to the waveform to locate action potentials. A primitive choice is a threshold detector, but more complex detectors make use of the detailed shape of the waveform. In this way, an analog voltage signal is processed into a binary response vector ρ_i that contains the spike train of the neuron. Its components are either zero or one, indicating whether or not a spike occurred during each 0.5 ms time bin.

5   Firing rate

In Figure 1 the stimulus amplitude is plotted versus time. Overlaid on top of this graph is the spike train. Inspection of this figure reveals that the frequency of action potential firing increases when the stimulus increases.

To quantify this relationship, it is necessary to transform the binary vector ρ_i into a firing rate or related quantity. For example, one can compute the number of spikes in a
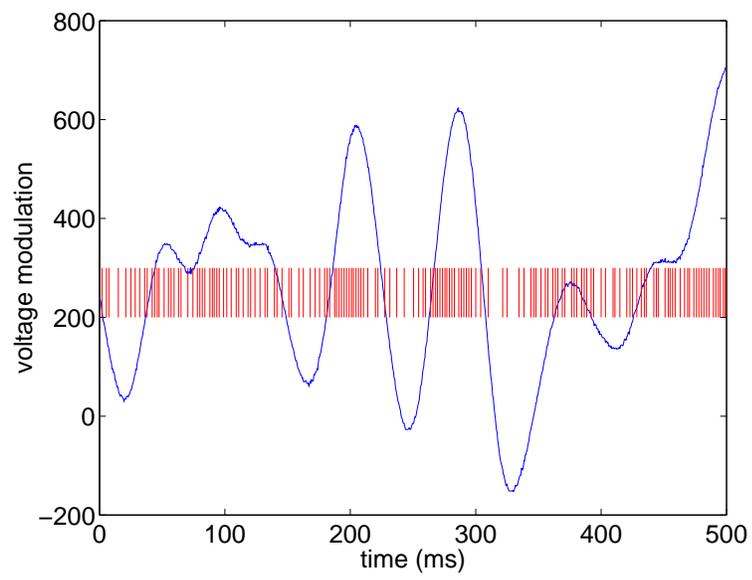
Figure 1: Spike train and stimulus amplitude versus time.

window of length $2M + 1$ centered at i.

$$n_i = \sum_{j=-M}^{M} \rho_{i+j}$$

This is like dragging a window across the spike train, and keeping track of the number of spikes in the window at each time.

Dividing n_i by the length of the window $2M + 1$ yields the probability of firing

$$p_i = \frac{1}{2M + 1} \sum_{j=-M}^{M} \rho_{i+j}$$

per time bin. Alternatively, one can define the firing rate

$$\nu_i = \frac{1}{(2M + 1)\Delta t} \sum_{j=-M}^{M} \rho_{i+j}$$

where Δt is the sampling interval. Assuming that the sampling interval is given in seconds, this gives rate in spikes per second, or Hz.

The above computations can be implemented using the `conv` operation in MAT-LAB, as will be explained in a later lecture and in the homework assignment.

Note that there are many other definitions of firing rate. This is just a simple one.

6 A linear model

If the pairs (s_i, p_i) are plotted as points on a graph, a linear relationship can be seen. The slope and intercept of the line can be found by optimizing the approximation $p_i \approx a + bs_i$ with respect to the parameters a and b.

So in this case, the neural coding problem can be addressed by simply fitting a straight line to data points. This is probably the most common way to fit experimental data in all of the sciences. Before we describe the technique below, let's pause to note that this is a very simple dataset. The stimulus is a scalar signal that varies with time. More generally, a vector might be required to describe the stimulus at a given time, as in the case of a dynamically varying image. The neural response might also be more complicated, if the experiment involved simultaneous recording of many neurons. But even in these more complex cases, it is sometimes possible to construct a linear model. When we do so later, we will see that some of the simple concepts introduced below can be generalized.

7 Fitting a straight line to data points

Suppose that we are given measurements (x_i, y_i), where the index i runs from 1 to m. In the context of the previous experiment, the measurements are (s_i, p_i). We have

simply switched notation to emphasize the generality of the problem. Our task is to find parameters a and b so that the approximation

$$y_i \approx a + bx_i \tag{1}$$

is as accurate as possible. Note that it is not generally possible to find a and b so that the error vanishes completely. There are two reasons for this. First, measurement are not exact, but suffer from experimental error. Second, while linear models are often used in computational neuroscience, the underlying behavior is not truly linear. The linear model is just an approximation. Note that this is unlike the case of physics, where the proportionality of force and acceleration ($F = ma$) is considered a true "law."

While there are many ways of finding an optimal a and b, the canonical one is the method of least squares. Its starting point is the squared error function

$$E = \sum_{i=1}^{m} \frac{1}{2}(a + bx_i - y_i)^2 \tag{2}$$

which quantifies the accuracy of the model in Eq. (1). If $E = 0$ the model is perfect. Minimizing E with respect to a and b is a reasonable way of finding the best approximation. Since E is quadratic in a and b, its minimum can be found by setting the partial derivatives with respect to a and b equal to zero.

Setting $\partial E / \partial a = 0$ yields

$$0 = ma + b\sum_i x_i - \sum_i y_i$$

while setting $\partial E / \partial b = 0$ produces

$$0 = \sum_i (a + bx_i - y_i)x_i \tag{3}$$

$$= a\sum_i x_i + b\sum_i x_i^2 - \sum_i y_i x_i \tag{4}$$

Rearranging slightly, we obtain two simultaneous linear equations in two unknowns,

$$ma + b\sum_i x_i = \sum_i y_i \tag{5}$$

$$a\sum_i x_i + b\sum_i x_i^2 = \sum_i y_i x_i \tag{6}$$

As a shorthand for the coefficients of these linear equations, it is helpful to define

$$\langle x \rangle = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \langle x^2 \rangle = \frac{1}{m}\sum_{i=1}^{m} x_i^2 \tag{7}$$

$$\langle y \rangle = \frac{1}{m}\sum_{i=1}^{m} y_i \qquad \langle xy \rangle = \frac{1}{m}\sum_{i=1}^{m} x_i y_i \tag{8}$$

The quantity $\langle x \rangle$ is known as the mean or first moment of x, while $\langle x^2 \rangle$ is known as the second moment. The quantity $\langle xy \rangle$ is called the correlation of x and y.

With this new notation, the equations for a and b take the compact form

$$a + b\langle x \rangle = \langle y \rangle \tag{9}$$
$$a\langle x \rangle + b\langle x^2 \rangle = \langle xy \rangle \tag{10}$$

We can solve for a in terms of b via

$$a = \langle y \rangle - b\langle x \rangle \tag{11}$$

This can be used to eliminate a completely, yielding

$$b = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\langle x^2 \rangle - \langle x \rangle^2} \tag{12}$$

Backsubstituting this expression in Eq. (11) allows us to solve for a.

The numerator and denominator in Eq. (12) have special names. The denominator $\langle x^2 \rangle - \langle x \rangle^2$ is called the variance of x, because it measures how much x fluctuates. Note that if all the x_i are equal to a large constant C, the second moment $\langle x^2 \rangle = C^2$ is large also. In contrast, the variance vanishes completely. The meaning of the variance is also evident in the identity

$$\langle (\delta x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$$

which you should verify for yourself. This equation says that the variance is the second moment of $\delta x = x - \langle x \rangle$, which is the deviation of x from its mean. The standard deviation is another term that you should learn. It is defined as the square root of the variance.

The numerator $\langle xy \rangle - \langle x \rangle \langle y \rangle$ in Eq. (12) is called the covariance of x and y. It is equal to the correlation of the fluctuations δx and δy,

$$\langle \delta x \delta y \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

Again, I recommend that you verify this identity on your own.

In summary, we have a simple recipe for a linear fit. Compute the covariance $\mathrm{Cov}(x, y)$ of x and y, and the variance $\mathrm{Var}(x)$ of x. The ratio of these two quantities gives the slope b of the linear fit. Then compute a by Eq. (11).

Substituting Eq. (11) in the linear approximation of Eq. (1) yields

$$y_i - \langle y \rangle \approx b(x_i - \langle x \rangle)$$

In other words, the constant a is unnecessary, if the linear fit is done to δx and δy, rather than to x and y. Given this fact, one approach is to compute the means $\langle x \rangle$ and $\langle y \rangle$ first, and subtract them from the data to get δx and δy. Then apply the formula

$$b = \frac{\langle \delta x \delta y \rangle}{\langle (\delta x)^2 \rangle}$$

which is equivalent to Eq. (12). The trick of subtracting the mean comes up over and over again in linear modeling.

Some of you may already have encountered the correlation coefficient r, which is defined by

$$r = \frac{\langle xy \rangle - \langle x \rangle \langle y \rangle}{\sqrt{\langle x^2 \rangle - \langle x \rangle^2}\sqrt{\langle y^2 \rangle - \langle y \rangle^2}}$$

You may have learned that r close to ± 1 means that the linear approximation is a good one. The correlation coefficient is similar to the covariance, except for the presence of the standard deviations of x and y in the denominator. The denominator normalizes the correlation coefficient, so that it must lie between -1 and 1, unlike the covariance, which can take on any value in principle. If you know the Cauchy-Schwarz inequality, you can use it to prove that $-1 \leq r \leq 1$, but this is not so illuminating.

The correlation coefficient can be interpreted as measuring the reduction in variance that comes from taking a linear (first-order) model of the data, as opposed to a constant (zeroth-order) model. Recall that the squared error of Eq. (2) measures the variance of the deviation of the data points from the straight line. This variance vanishes only when the model is perfect.

For the best zeroth-order model, we constrain $b = 0$ in Eq. (2), so that E is minimized when $a = \langle y \rangle$, taking a value proportional to the variance of y. For the best first-order model, E is minimized with respect to both a and b, so that its optimal value is further reduced. The ratio of the new E to the old E is $1 - r^2$. Another way of saying it is that r^2 is the fraction of the variance in y that is explained by the linear term in the model.