Neural dynamics of vision

A computational perspective

CLAYTON SEITZ[1]

February 4, 2021

---

[1] `cwseitz.github.io`

Dedicated to Calvin and Hobbes.

Contents

1 The Neural Code **1**
- 1.1 Introduction 1
- 1.2 Neuroelectronics 1
  - 1.2.1 RC Model 2
  - 1.2.2 Equilibrium and Reversal Potentials 3
  - 1.2.3 Membrane Current 4

2 Gabor Theory and Image Statistics **7**
- 2.1 Section heading 7

3 Information and Coding Theory **9**
- 3.1 Introduction 9
- 3.2 Entropy 9
  - 3.2.1 Joint and Conditional Entropy 10
- 3.3 KL-Divergence and Mutual Information 11
  - 3.3.1 The Data-Processing Inequality 12
- 3.4 Source Coding 12
  - 3.4.1 Kraft's Inequality 13
  - 3.4.2 Source Coding Theorem 13
  - 3.4.3 Fano's Inequality 13
  - 3.4.4 Jensen's Inequality 14
  - 3.4.5 Fano's Inequality 14
- 3.5 Error Correcting Codes 14

4 Microscopy and Image Analysis **15**
- 4.1 Section heading 15

1

The Neural Code

1.1 Introduction

Throughout this section of the text we will coarsely examine the biophysical properties of the elements of the nervous systems: neurons. The hope is that by introducing the nervous system from a physical perspective, much of the mathematical developments in later chapters will seem natural and well-motivated. However, we will see that much of the later work neglects many of the features of neurons this chapter will introduce and for good reason. It never hurts to know what you have left out of your model.

Neurons are one of the most interesting cell types in animals particularly for their evolved ability to sense their external environment and produce a myriad of responses in favor of the survival of the host organism. All vertebrates have a distinct central nervous system consisting of a brain and a spinal cord which process information from or send information to a peripheral nervous system. Here we will discuss the cellular elements that allow neurons to send information to each other in such a system.

1.2 Neuroelectronics

Neurons are bounded by a 3-4nm thick membrane composed of a diverse set of lipids forming a bi-layer and are submerged in a bath of water and ions *in vivo*. These cells have evolved a vast array of membrane ion channels that control the concentrations of ions on either side of the membrane. The degree to which these channels allow ions to flow in and out of the membrane is the basis of all neural processing and in turn all of our actions and conscious awareness.

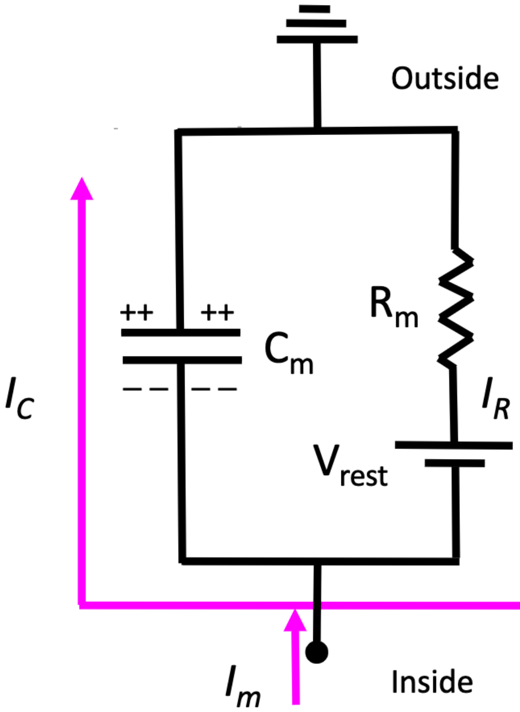

Figure 1.1: RC Circuit Model

1.2.1 RC Model

A well-known model of a neuron membrane is an RC circuit where the resistor represents a sort of permeability of the membrane to ions while the capacitor models the membrane itself. Since either side of the membrane contains electrically charged ions there exists an electrical potential in these regions and in turn a potential difference across the lipid bi-layer. The current through the membrane is ultimately determined by the balance between diffusion of ions down their concentration gradients and electrical potential gradients.

If we inject a current I_E into the cell, by conservation of current we have

$$I_E = C\frac{dV}{dt} + I_R$$

where V is the membrane voltage. We can also use Maxwell's equations (Kirchhoff's voltage rule) to write

$$I_R = \frac{V - V_0}{R}$$

where have replaced the resistance by the conductance g. Combining these two equations we are left with the following differential equation for the membrane voltage.

$$C\frac{dV}{dt} = I_E - \frac{V - V_0}{R}$$

which is commonly rewritten by defining the membrane time constant $\tau = RC$

$$\tau\frac{dV_C}{dt} = I_E R - (V - V_0)$$

Notice that as $t \to \infty$ we have $V_C - V_{rest} \to I_E R + V_0$ and so we define the steady-state voltage $V_\infty = I_E R + V_0$. Inserting this new definition into the equation

$$\tau\frac{dV}{dt} = V_\infty - V$$

which can be solved to give

$$V(t) = V_\infty(1 - e^{-\frac{t}{\tau}})$$

1.2.2 Equilibrium and Reversal Potentials

The flux of ions through the cell membrane is due to an imbalance of electrical and chemical potentials. Recall the Boltzmann distribution of energies at a temperature T

$$P(\epsilon) = \frac{e^{-\frac{\epsilon}{k_B T}}}{Z}$$

where the partition function $Z = \int e^{-\frac{\epsilon}{k_B T}} d\epsilon$. We define the **reversal potential** as the membrane potential at which electrical forces perfectly balance the diffusion of ions down their concentration gradient and their is no current through the membrane. We compute this using the famous Nernst equation

$$E = \frac{RT}{zF}\ln\frac{[o]}{[i]}$$

Perhaps we would rather compute the resting potential of the membrane with knowledge of the concentrations of ions inside and outside of the cell. For a set of positive ions M and negative ions A

$$E = \frac{RT}{F} \ln \frac{\sum_i M_i[M_i]_{out} + \sum_i A_i[A_i]_{in}}{\sum_i M_i[M_i]_{in} + \sum_i A_i[M_i]_{out}}$$

where we have used M_i, A_i to represent the permeability of these ions.

1.2.3 Membrane Current

We saw in our basic RC circuit model of the cell membrane that the membrane current I_R was very simply given by Ohm's law $I_R = (V - V_0)/R$. This very simple model assumes that the entire membrane can be lumped as a single resistor R; however, in reality, neural membranes have a diverse set of channel types each with their own respective resistance to current flow. Therefore, we need to expand R into a set of parallel resistors each with their own resistance.

$$I = \sum_i g_i(V - E_i)$$

This setup is more accurate but still not entirely correct; Ohm's law assumes that we are dealing with a **linear circuit** where the resistance is a constant (not a function of the voltage). In the cell, many channels are **voltage gated** which means that they *are* a function of the voltage and Ohm's law doesn't apply. A more correct definition of the total membrane current is then

$$I = \sum_i g_i(V)(V - E_i) \tag{1.1}$$

The functions $g_i(V)$ are difficult to estimate due to the complexity of ion channels and their mechanisms of opening and closing. Hodgkin and Huxley came up with a model for these voltage-dependent conductance for sodium and potassium. Their model defines what we call an **open probability** for a single ion channel at a particular voltage. In the case of the potassium channel, Hodgkin and Huxley thought of the channel as a structure composed of 4 particles which each probability n of being in the *permissive state*. If all four particles are in such a state, the channel will open. Each particle has a probability of being in this state which increases with the voltage sigmoidally

$$n(V) = \frac{1}{1 + (A_\beta/A_\alpha)\exp((B_\alpha - B_\beta))V/V_T} \tag{1.2}$$

and so the total probability the channel is open is n^4. The model of the sodium channel is slightly different. They introduced two probabilities

m and h where m is directly analogous to n and h is the probability that a gate on the sodium channel is open. They fit their experimental data using m^3 and so probability the sodium channel is open is then m^3h. The total membrane current according to the Hodgkin-Huxley model can then be found by expanding Eq. 1.1

$$I = \bar{g}_{Na}m^3h(V - E_{Na}) + \bar{g}_K n^4(V - E_K) + \bar{g}_L(V - E_L) \tag{1.3}$$

where a leakage conductance g_L is also included.

2

Gabor Theory and Image Statistics

Here is a chapter

2.1 Section heading

3

Information and Coding Theory

> *"We may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it"*
>
> – Claude Shannon

3.1 Introduction

Information theory is a framework first introduced by Claude Shannon's seminal paper *A mathematical theory of communication* published in 1948. At it's core, information theory makes the intuitive concept of *information* mathematically rigorous and forms the foundation of many modern communication systems. Neural circuits in the visual system are an especially interesting example of such a communication system. Therefore, in this section, the information theoretic concepts necessary for studying neural circuits are introduced.

3.2 Entropy

The concept of entropy is not exclusive to information theory; rather, it is used widely in disciplines such as physics and mathematical statistics. In fact, entropy was originally defined in statistical physics when Ludwig Boltzmann gave a statistical description of a thermodynamic system of particles. Since this is arguably the more intuitive path as opposed an entirely mathematical description, I will follow a similar line of reasoning in the following paragraphs.

In every application, the entropy $\mathbf{H}$ is a measure of uncertainty or how much information is contained in a random variable x. In information theory, the entropy is a property of a probability distribution of a random variable

$P(x)$ where x can take on continuous or discrete values. For the discrete case, we can express the entropy in bits

$$\mathbf{H} = \sum_{x \in S} P(x) \log \frac{1}{P(x)} \tag{3.1}$$

where the set S spans the entire space of possible discrete values of x. We can go on to derive upper and lower bounds for the entropy. Notice that $\mathbf{H} \geq 0$ since $P(x) \leq 1$ and therefore $\log P(x) \leq 0$ for all x. At the same time, if we define a variable $Y = \frac{1}{\log x}$, we can write

$$\begin{aligned}
\mathbf{H} &= \mathbf{E}[\log Y] \\
&\leq \log \mathbf{E}[Y] \\
&= \log \sum_{y} P(x) \frac{1}{P(x)} \\
&= \log |S|
\end{aligned}$$

which is just the entropy of a uniform distribution.

3.2.1 Joint and Conditional Entropy

In this section, we discuss joint and conditional entropy which are really just two sides of the same coin

$$\begin{aligned}
\mathbf{H}(X,Y) &= \sum_{x,y} P(x,y) \log \frac{1}{P(x,y)} \\
&= \sum_{x,y} P(x)P(y|x) \log \frac{1}{P(x)P(y|x)} \\
&= \sum_{x,y} P(x)P(y|x) \log \frac{1}{P(x)} + \sum_{x,y} P(x)P(y|x) \log \frac{1}{P(y|x)} \\
&= \sum_{x,y} P(x)P(y|x) \log \frac{1}{P(x)} + \sum_{x} P(x) \sum_{y} P(y|x) \log \frac{1}{P(y|x)} \\
&= H(X) + H(Y|X)
\end{aligned}$$

This result defines the **chain rule** for entropy. We typically refer to the term $H(Y|X)$ as the **conditional entropy**. It can be calculated independently using the following definition

$$\begin{aligned}
H(X|Y) &= \mathbf{E}_y H(X|Y=y) \\
&= \mathbf{E}_y \sum_{x} P(X|Y=y) \log \frac{1}{P(X|Y=y)}
\end{aligned}$$

Furthermore, it can be shown that the chain rule derived above applies to a tuple of random variables longer than two.

$$H(X_1, \dots, X_m) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \dots H(X_m|X_1 \dots X_{m-1})$$

Recalling that conditioning reduces entropy or does nothing at all, we can write

$$H(X_1, \dots, X_m) \leq H(X_1) + H(X_2) + \ \dots \ + H(X_m)$$

which is referred to as the **subadditivity** property of entropy. We should also address what to do when we need to compute the entropy of a joint distribution (X, Y) conditioned on a variable Z or when Z itself is conditioned on a joint distribution. These two things are related by using the chain rule for joint entropy

$$H(X, Y|Z) \quad = \quad H(X, Y) + H(Z|X, Y)$$

Now we will prove that conditioning the distribution of a random variable X on another variable Y i.e. can reduce the entropy of X. What we need to show is that $H(X|Y) - H(X) \leq 0$.

3.3 KL-Divergence and Mutual Information

The Kullbeck-Leiber distance or **KL Divergence** is a measure of the distance between two distributions over a random variable X. Assume we have two distributions P, Q on a random variable X where P is the correct distribution on X and Q is an incorrect distribution. By definition, the KL-Divergence $D_{KL}(P||Q)$ is the extra information (bits) it takes to communicate X when using the incorrect distribution Q. To be precise, $H(Q) = H(P) + D_{KL}(P||Q)$.

Definition 1. *The KL-Divergence is*

$$D_{KL}(P||Q) = \sum_X P(X) \log \frac{P(X)}{Q(X)}$$

Additionally, we can show that KL-Divergence follows a chain rule

Furthermore, an indispensable tool in information theory is the idea of **mutual information** which, as the name suggests, measures the amount of overlapping information in a pair of random variables. More formally, it is the KL-Divergence between the joint distribution of the pair of variables and the product of their marginal distributions (which implies they are independent)

Definition 2. *The mutual information is*

$$\begin{aligned} I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y)) \\ &= \sum_x \sum_y P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \end{aligned}$$

A very useful property of the mutual information is that it is strongly related to conditional entropy and statistical independence. Conditional entropy tells us how much information is contained in a variable X which its distribution is conditioned on Y. We might expect that this conditioning doesn't really have an effect if X and Y are completely independent. Indeed,

$$\begin{aligned} I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y)) \\ &= \sum P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= \sum P(x) \log \frac{1}{P(x)} + \sum P(x,y) \log \frac{P(x,y)}{P(y)} \\ &= H(X) - H(X|Y) \end{aligned}$$

Note that this result implies that $I(X;Y) = I(Y;X)$. We will next address the mutual information between a distribution on X and a joint distribution (Y,Z) making use of the relationship derived above.

$$\begin{aligned} I(X;(Y,Z)) &= H(X) - H(X|Y,Z) \\ &= H(X) + H(Y,Z|X) - H(Y,Z) \end{aligned}$$

Finally, we look at the mutual information between a distribution on X and a conditional distribution $Y|Z$.

3.3.1 The Data-Processing Inequality

The data-processing inequality states that if a function operates on a random variable X it can only decrease its entropy. That is, for any function f s.t. $Y = f(X)$, we have that $H(Y) \geq H(X)$. We can prove that this is true using the mutual information $I(X;Y)$.

3.4 Source Coding

Definition 3. *A code of a set S that uses an alphabet Ω is a map $C : S \to \Omega$ that assigns each element of S a finite string over the alphabet Ω. We say that the mapping C is* **prefix free** *if for all pairs $x, y \in S$ where $x \neq y$, $C(x)$ is not a prefix of $C(y)$.*

Most of the time the alphabet Ω we use is the set $0, 1$.

3.4.1 Kraft's Inequality

Definition 4. *For a binary code, there exists a prefix free code C with codeword lengths l_i if and only if*

$$\sum_i 2^{-l_i} \leq 1 \tag{3.2}$$

At this point we would like to apply the concept of entropy to source coding. Indeed, it is true that if we have a random variable X over the set S, the minimum number of bits it will take us to communicate the value of X on average is the entropy $H(X)$.

Proof. The expected number of bits to communicate X is given by $\sum_x p(x)|C(x)|$

$$\begin{aligned} H(X) - \sum_x P(x)|C(x)| &= \sum P(x)[\log \frac{1}{P(x)} - |C(x)|] \\ &= \sum P(x) \log \frac{1}{P(x)2^{|C(x)|}} \\ &\geq \log \sum P(x) \frac{1}{P(x)2^{|C(x)|}} \\ &= \log \sum \frac{1}{2^{|C(x)|}} \\ &\leq 0 \end{aligned}$$

□

by Kraft's inequality for prefix-free codes.

3.4.2 Source Coding Theorem

So far we have seen how to construct a prefix-free code and that the absolute lower bound on the number of bits it takes to encode a random variable is its entropy. Next, we would like to answer the following question: how do we actually design a code to communicate a random variable X so that it approaches this lower bound? The answer is addressed by the *fundamental source coding theorem*

Theorem 1. *For all epsilon > 0 there is a $n_0 \leq n$ such that given n instances of a variable X it is possible to communicate X with $H(X) + \epsilon$ bits on average.*

This means that we can approach the entropy by increasing n.

3.4.3 Fano's Inequality

Fano's inequality quantifies the amount of information lost in a noisy channel given some probability of transmission error.

3.4.4 Jensen's Inequality

Jensen's inequality is a statement about convexity. Consider a binary variable x that takes the value 0 with probability α and value 1 with probability $1 - \alpha$.

$$x = \begin{cases} 0 & \alpha \\ 1 & 1 - \alpha \end{cases}$$

A function f of the variable x is said to be *convex* if the following inequality holds

$$\alpha f(x) + (1 - \alpha) f(y) \leq f(\alpha x + (1 - \alpha) y)$$

which when generalized for an arbitrary random variable x forms Jensen's inequality

$$\mathbf{E}[f(x)] \leq f(\mathbf{E}[x]) \tag{3.3}$$

Example

Let's consider a function $f : \mathbb{R} \to \mathbb{R}$. Using Jensen's inequality, we can prove that $f = x^2$ or $f = x \log x$ are convex functions. Let's begin by applying it to x^2 for a general normalized probability distribution $p(x)$.

$$\begin{aligned} \int p(x) f(x) dx &= \int x^2 p(x) dx \\ &= x^2 - 2 \int x dx \\ &= 0 \leq x^2 \; \forall x \end{aligned}$$

We have a similar proof for $f(x) = x \log x$

$$\begin{aligned} \int p(x) f(x) dx &= \int x \log x \; p(x) dx \\ &= x \log x - \int \frac{d}{dx} x \log x \; dx \\ &= 0 \leq \mu \log \mu \end{aligned}$$

where $\mu = \mathbf{E}[x] \geq 0$ since f is only defined on $[0, \infty]$.

3.4.5 Fano's Inequality

3.5 Error Correcting Codes

4

Microscopy and Image Analysis

"This is a quote and I don't know who said this."
– Author's name, *Source of this quote*

4.1 Section heading