# Gatsby Theoretical Neuroscience Notes

Jorge A. Menendez

January 26, 2020

## Contents

# 1 Notes

D&A stands for Dayan & Abbott *Theoretical Neuroscience* textbook ([Dayan and Abbott, 2001]).
ND stands for *Neuronal Dynamics* textbook by Gerstner, Kistler, Naud, & Paninski ([Gerstner et al., 2014]).

# 2 Biophysics

## 2.1 Single-Compartment Models (Soma)

We treat the neuron as a circuit (see appendix 7.3), with

- *Membrane current* $i_m$ (amps per unit area), the outward/inward flow of positive/negative ions
- *Membrane potential* $V$ (millivolts), the voltage difference between outside and inside of the cell
- *Specific membrane capacitance* $c_m$ (farads per unit area), an inherent property of the phospholipid bilayer
- *Total membrane resistance* $R_m$ (ohms), depends on the density and types of ion channels on the membrane. $R_m$ is inversely proportional to area $A$, via the *specific membrane resistance* $r_m$ (ohms-unit area), which is the total membrane resistance when $A = 1$:

$$R_m = \frac{r_m}{A}$$

- *Specific conductance* $g_i$ (siemens per unit area), the inverse resistance of the membrane to ions $i$, depending on the particular ion channel
- External injected current $I_e(t)$ (amps)

Importantly, neurons actively maintain a concentration gradient with respect to the extracellular space, generating diffusion of particular ions. Two crucial ones are potassium ($K^+$) and sodium ($Na^+$) ions, which are respectively pumped in and out of the cell via the active sodium-pottasium pump on the membrane to maintain a relatively higher/lower concentration of $K^+/Na^+$ inside than outside the cell. By pumping out 3 $Na^+$ ions for every 2 $K^+$ pumped in, this pump (along with others) creates an equilibrium *resting* membrane potential at about $-70$mV.

Note that this is a small enough potential such that ion flow is affected both by diffusion and electrical forces (i.e. thermal energy $\approx$ potential energy, see D&A pg. 155). Since the membrane current depends on ion movement, this means we need to slightly modify our usual electric circuit equations. Specifically, to get an expression for the membrane current produced by the flow of type $j$ ions, we modify Ohm's Law to incorporate the ionic flow generated by diffusion:

$$i^{(j)} = g_j(V - E_j)$$

where $E_j$ is called the *reversal potential* of ion $j$: the membrane potential necessary to counteract the flow produced by diffusion such that there is 0 net ion flow (i.e. 0 current). For ions that are pushed into the cell by diffusion (i.e. higher those with higher extracellular than intracellular concentration), their reversal potential will be of the same polarity as their charge so as to repel them from the membrane and balance out the diffusion forces. For example, we need a highly positive membrane potential to repel the positively charged sodium ions against their concentration gradient. And vice-versa: we need a highly negative membrane potential to attract potassium ions to move against their concentration gradient. In fact we have

| Ion | Concentration Gradient | Reversal Potential $E_i$ |
|---|---|---|
| $K^+$ | higher inside | $\sim -75$mV |
| $Cl^-$ | higher outside | $\sim -65$mV |
| $Na^+$ | higher outside | $\sim 50$mV |
| $Ca^{2+}$ | higher outside | $\sim 150$mV |

We thus call $Na^+$ and $Ca^{2+}$ conductances *depolarizing* since they push the membrane potential above zero, whereas $K^+$ and $Cl^-$ conductances are *hyperpolarizing*. But it is important to note that this classification hinges on the equilibrium potentials, which depend on the intra/extracellular ion concentration gradients. In the below, we will always assume these fixed, but in reality they may change dynamically on long timescales. For example, intra-cellular concentrations of $Cl^-$ increase during development so that the $Cl^-$ conductance becomes depolarizing (effectively turning GABAergic synapses into excitatory synapses).

Note that by our above equation, current is positive when the membrane potential is above the reversal potential, which is when cations (positively charged ions) flow out and anions (negatively charged ions) flow in to the cell. So the membrane current can be thought of as the total *outflow* of cations from or *inflow* of anions to the cell, which we can write as:

$$i_m = \sum_j g_j(V - E_j)$$

where we have summed over all ionic conductances $j$. Importantly, the conductances $g_j$ may depend on other factors. Below, we consider constant (*passive*) conductances and voltage-dependent (*active*) conductances. In the single-compartment models we consider, we work in the approximation of the soma as having spatially uniform electric properties, such that the *total membrane capacitance* is given by $C_m = Ac_m$, where $A$ is the total surface area of the cell and $c_m$ is assumed constant.

Using our equation for RC circuits (section 7.3), we then have dynamics

$$c_m \frac{\mathrm{d}V}{\mathrm{d}t} = -i_m + \frac{I_e(t)}{A}$$

where we have noted that the rate of change in membrane potential is equal to the rate of inflow of positive ions (= *negative* membrane current, see end of previous paragraph) + injected external current (which needs to be converted to current per unit area). Multiplying both sides now by the specific membrane resistance $r_m$, we have

$$\tau_m \frac{\mathrm{d}V}{\mathrm{d}t} = -r_m i_m + V_e(t)$$

where $V_e(t) = \frac{r_m I_e(t)}{A} = R_m I_e(t)$ and $\tau_m = r_m c_m$. $\tau_m \sim 10 - 100$ms is called the membrane time constant, which sets the timescale for changes in membrane potential. Note that it is independent of the cell surface area ($\tau_m = R_m C_m = \frac{r_m}{A} \times Ac_m = r_m c_m$).

### 2.1.1 Passive Protein Channels

We first consider only those currents generated by the ion pumps maintaining the ionic concentration gradient across the neuron cell membrane. We can generally assume that these work at a relatively constant rate, such that we can safely make the approximation $g_j \approx \bar{g}_j$, where the bar indicates $\bar{g}_j$ is constant. Summing together all such channels $j$, we can express this so-called *leak current* by

$$i_L(t) = \sum_j \bar{g}_j(V(t) - E_j) = \bar{g}_L(V(t) - E_L)$$

where $\bar{g}_L = \sum_j \bar{g}_j$ and $E_L = \frac{\sum \bar{g}_j E_j}{\sum \bar{g}_j}$, and I have explicitly written the time-dependence of the dynamic variables. Setting $i_m = i_L$ gives us the classic *leaky integrate-and-fire* (LIF) model neuron

$$\tau_m \frac{\mathrm{d}V}{\mathrm{d}t} = -(V - E_L) + V_e(t)$$

since $r_m = \frac{1}{\bar{g}_L}$ in the absence of any other currents. In this model, when the membrane potential $V$ reaches the spiking threshold $V_{th} \approx 50$mV, the neuron emits a spike and $V$ is reset to some $V_0$. In the absence of external input, the membrane potential decays exponentially to $E_L$, which is typically set to the resting membrane potential.

Using the substitution $u(t) = V(t) - E_L$, we can quickly solve this using the method of integrating factors (section 7.4.1), giving us

$$V(t) = E_L + \frac{1}{\tau_m} \int_0^t V_e(t') e^{\frac{-(t-t')}{\tau_m}} \, \mathrm{d}t'$$

where we dropped the additive constant corresponding to the initial condition by assuming a large $t$. This expression is easily interpreted as demonstrating the characteristic memorylessness of LIF neurons: whatever external input $V_e(t')$ is received at time $t'$, its effect on the neuron membrane potential $V(t' + \Delta t)$ decays exponentially as $\Delta t$ grows, with time constant $\tau_m$.

If $I_e(t) = I_e$ is constant (so $V_e(t) = V_e = R_m I_e$), it turns out we can do a little more analytically. In this case, the exact solution to the differential equation becomes

$$V(t) - E_L = V_e + (V(0) - E_L - V_e)e^{\frac{-t}{\tau_m}}$$

Let the neuron spike whenever $V(t)$ reaches the threshold $V_{th}$, then immediately resetting to $V_0$ with no refractory period. Given that the initial condition is the reset potential $V(0) = V_0$, we can compute the *interspike interval* $t_{isi}$

$$V(t_{isi}) = E_L + V_e + (V_0 - E_L - V_e)e^{\frac{-t_{isi}}{\tau_m}} = V_{th}$$
$$\Leftrightarrow t_{isi} = \tau_m \log \frac{V_e - (V_0 - E_L)}{V_e - (V_{th} - E_L)}$$

giving us the *interspike interval firing rate*

$$r_{isi} = \frac{1}{t_{isi}} = \frac{1}{\tau_m} \log \left(1 + \frac{V_{th} - V_0}{V_e - (V_{th} - E_L)}\right)^{-1} \approx \left[\frac{V_e + E_L - V_{th}}{\tau_m (V_{th} - V_0)}\right]_+$$

where the approximation holds for large $I_e$, since $\log(1 + z) \approx z$ for small $z$, and $[]_+$ is linear rectification. In this regime, the firing rate is linear in $I_e = V_e/R_m$.

In fact, this turns out to be a pretty good approximation for short periods of stimulation (D&A fig 5.6). For longer periods of constant current injection, however, the ISI lengthens over time, a phenomenon called *spike-rate adaptation*. This is easily incorporated into the LIF neuron by adding a time-dependent hyperpolarizing potassium channel conductance that decays exponentially in time:

$$\tau_m \frac{\mathrm{d}V}{\mathrm{d}t} = -r_m i_L - r_m g_{sra}(t)(V - E_K) + V_e(t)$$
$$\tau_{sra} \frac{\mathrm{d}g_{sra}}{\mathrm{d}t} = -g_{sra} + \tau_{sra} \Delta_{sra} \sum_k \delta(t - t^{(k)})$$

where $t^{(k)}$ is the time of the $k$th spike and $\Delta_{sra}$ is the rise in the hyperpolarizing conductance at the time of a spike (for this to be the case, it is necessary to include the $\tau_m$ constant in front of the $\delta$ functions). In this model, $g_{sra}$ instantly jumps up by $\Delta_{sra}$ whenever the neuron spikes, leading to hyperpolarizing current that make it more difficult to spike soon after.

### 2.1.2 Active Protein Channels

To model actual action potentials, we need to incorporate non-linearities via non-constant (i.e. *active*) membrane conductances $g_i(t)$. Specifically, we additionally model the opening and closing of particular ion channels, thus allowing the conductances of the ions they are permeable to to vary dynamically. The probability of a channel being open can depend on several factors, such as the membrane potential, the concentration or presence of neurotransmitters or neuromodulators, or other internal messengers such as $Ca^{2+}$. Importantly, it has been experimentally observed that the opening of ion channels tends to resemble a Poisson process (i.e. exponentially distributed inter-opening intervals), where the probability of the channel opening is independent of its past history (at a fixed membrane potential). Thus, the opening of any given type $i$ channel on the membrane can be treated as an i.i.d. Bernoulli random variable, so that, in the limit of many channels, the law of large numbers tells us that the proportion of channels open should be $\approx$ the probability of a channel being open. We can thus model the specific conductance of an ion $i$ as

$$\begin{aligned} g_i(t) &= \bar{g}_i^{open} \times \text{channel } i \text{ density} \times \text{proportion of } i\text{-channels open at time } t \\ &= \bar{g}_i P(\text{channel } i \text{ is open at time } t) \end{aligned}$$

where $\bar{g}_i^{open}$ is the conductance of the channel when it is open, such that $\bar{g}_i$ is the channel $i$ conductance per unit area - a constant I will call the *maximal conductance* of channel $i$ (a constant).

We then model the open probability as time-varying and dependent on some other factor (e.g. membrane potential, neurotransmitter concentration, etc.).

We consider here the Hodgkin-Huxley model, which models the opening and closing of *voltage-dependent* $K^+$ and $Na^+$ channels. More specifically, we model the opening and closing probabilities of different "gates" of each channel, the probability of opening or closing being dependent on the membrane potential. For a fixed membrane potential $V$, we thus model the opening and closing of a given gate $m$ as a two-state Markov model with states $\{1, 2\} = \{open, closed\}$ with transition matrix

$$A^{(m)} = \begin{bmatrix} 1 - \beta_m(V) & \alpha_m(V) \\ \beta_m(V) & 1 - \alpha_m(V) \end{bmatrix}$$

where $A^{(m)}_{ij}$ is the probability of transitioning from state $j$ to state $i$, where $\alpha_m(V), \beta_m(V)$ are the probability per unit time (i.e. the rate) of gate $m$ opening or closing, respectively, at a fixed membrane potential $V$. For example, $A^{(m)}_{12}$ might be the probability per ms of gate $m$ closing, i.e. transitioning from state 1 (open) to state 2 (closed).

Consider a gate with probability of being open at time $t$ given by $m(t)$. Letting $P(\cdot)$ designate a probability proper and $\tilde{P}$ designating a probability per unit time (i.e. a rate), we can write down its dynamics as

$$\begin{aligned} \frac{\mathrm{d}m}{\mathrm{d}t} &= \tilde{P}(\text{transitioning to open}) - \tilde{P}(\text{transitioning to closed}) \\ &= P(\text{closed})\tilde{P}(\text{closed} \to \text{open}) - P(\text{open})\tilde{P}(\text{open} \to \text{closed}) \\ &= (1 - m)\alpha_m(V) - m\beta_m(V) \\ &= \alpha_m(V) - (\alpha_m(V) + \beta_m(V))m \end{aligned}$$

where we have the unique steady state

$$\frac{\mathrm{d}m}{\mathrm{d}t} = 0 \Leftrightarrow m_\infty(V) = \frac{\alpha_m(V)}{\alpha_m(V) + \beta_m(V)}$$

Recalling that $\alpha_m(V), \beta_m(V)$ are rates, we can write the dynamics as

$$\tau_m(V)\frac{\mathrm{d}m}{\mathrm{d}t} = m_\infty(V) - m$$

where the voltage-dependent time constant[1] for this gate is $\tau_m(V) = \frac{1}{\alpha_m(V)+\beta_m(V)}$. Thus, for a fixed membrane potential $V$, the probability of an $m$-gate being open exponentially approaches $m_\infty(V)$, with time constant $\tau_m(V)$.

In the Hodgkin-Huxley model, we assume that the $K^+$ and $Na^+$ channels have more than one "gate"[2]. The potassium ion channels have four such gates, such that the probability of a potassium channel being open is

$$P(K^+ \text{ open at time } t) = n^4(t)$$

with $\alpha_n(V), \beta_n(V)$ set such that $n_\infty(V)$ is a sigmoid function. We thus call the potassium conductance a *persistent conductance*, since it always increases with an increase in membrane potential (i.e. depolarization). Sodium ion conductance, on the other hand, is called a *transient conductance*, because in addition to three gates that open with depolarization, there is one "inactivation gate" that closes:

$$P(Na^+ \text{ open at time } t) = m^3(t)h(t)$$

with $m_\infty(V)$ a sigmoid and $h_\infty(V)$ a flipped sigmoid (see figure 1A). Thus, as the neuron becomes more depolarized and the $n$-gates open, the sodium ion channel is only transiently open before the $h$-gate closes.

The full Hodgkin-Huxley model is then (here, $\tau$ is the passive membrane time constant):

$$\begin{aligned} \tau\frac{\mathrm{d}V}{\mathrm{d}t} &= -(V - E_L) - \bar{\rho}_{\mathrm{K}} n^4(t)(V - E_{\mathrm{K}}) - \bar{\rho}_{\mathrm{Na}} m^3(t)h(t)(V - E_{\mathrm{Na}}) + V_e(t) \\ \tau_X(V)\frac{\mathrm{d}X}{\mathrm{d}t} &= X_\infty(V) - X, \qquad X = m, n, h \end{aligned}$$

[1] Not to be confused with the membrane time constant $\tau_m$!!! It just happens that the standard notation for the open probability of one of the Hodgkin-Huxley gates is $m$.

[2] The resulting exponents on each gating variable were originally fit to the data by Hodgkin and Huxley, but I believe it turns out that their interpretation as a number of charged gates turns out to be true (ref??)

where, recalling that both sides were multiplied by $r_m = \frac{1}{\bar{g}_L} = \frac{1}{\sum \bar{g}_i}$,

$$\bar{\rho}_Z = \frac{\bar{g}_Z}{\sum \bar{g}_i}, \quad Z = \text{Na, K}$$

with the sum being over passive conductances $i$. Importantly, $\tau_m(V) << \tau_h(V), \tau_n(V)$ for all reasonable $V$, so the $m$-gates open much faster than the $h$-gates close with a depolarization, allowing the sodium channels to remain open for a transient period (see fig. 1 below, D&A fig. 5.11). This transient period where sodium rushes into the cell and quickly depolarizes it is the action potential, immediately succeeded by a closing of the sodium channels (via the slow $h$-gate) and opening of the potassium channels (via the slow $n$-gates), thus letting potassium flow out of the cell (recall that potassium conductance is a hyperpolarizing conductance, $E_K < 0$) and hyperpolarizing the cell back to its resting membrane potential.

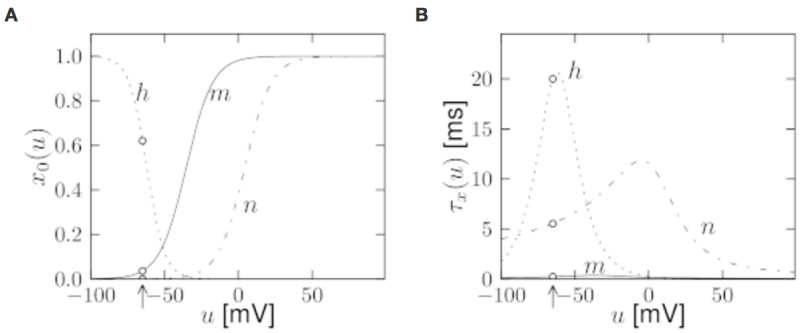



Figure 1: Copied from ND textbook figure 2.3. **A** plots $m_\infty(V), h_\infty(V), n_\infty(V)$, **B** plots $\tau_m(V), \tau_h(V), \tau_n(V)$, $u = V$ on abscissa. Arrow indicates resting potential $E_L = -65$mV. See section 2.2 for parameter settings - note that these plots are quantitatively quite different from the analagous plots in D&A figure 5.10. Not sure which one is empirically correct...

### 2.1.3 Reduced Single-Neuron Models: Switches & Type I/Type II Model Neurons

We can get a better understanding of this four-dimensional system of equations by reducing it to fewer dimensions. The most basic such approximation we can do is assume the dynamics of the gating variables to operate on a much faster timescale than the passive membrane potential dynamics (i.e. $\tau >> \tau_m, \tau_n, \tau_h$, which is approximately true for $m$-gates but not really for $n, h$) such that we can simply set the gating variables to their respective equilibria in the membrane potential dynamics, giving us the one-dimensional system:

$$\tau \frac{\mathrm{d}V}{\mathrm{d}t} = -(V - E_L) - \bar{\rho}_{\mathrm{K}} n_\infty^4(V)(V - E_{\mathrm{K}}) - \bar{\rho}_{\mathrm{Na}} m_\infty^3(V) h_\infty(V)(V - E_{\mathrm{Na}}) + V_e(t)$$

For appropriate parameter settings, at $V_e(t) = 0$ this system will have two stable fixed points at high and low $V$, separated by an unstable fixed point in between, i.e. a cubic function on the $V$ vs $\frac{\mathrm{d}V}{\mathrm{d}t}$ plane with three roots, a minima between the smaller root and the middle root, and a maxima between the middle and larger root. Changing $V_e(t)$ then simply shifts this cubic function up and down. Thus, if $V_e(t)$ increase above some threshold, the smaller two roots eventually disappear, leaving only the larger stable point. Conversely, if $V_e(t)$ decreases below some threshold, the larger two roots disappear and the system converges to the smaller stable point. We thus have a switch! This would be really useful for computation (think of how a digital computer works), but it has an unfeasible energetic cost: when the switch is at the ON state, the ion pumps need to work extra hard to keep the membrane potential at the larger stable point. Furthermore, actual membrane potential dynamics look nothing like this.

A more biologically realistic reduction is obtained by observing that (1) the approximation of $\tau_m(V) = 0$ is not a bad one and (2) $\tau_h(V)$ and $\tau_n(V)$ are on the same order over all $V$. We might then hope to replace $h(t), n(t)$ with an artificial variable $w(t)$ that can jointly represent their dynamics. Rigorously, we might do this by fitting a linear model to their dynamical coupling, e.g. $h(t) \approx an(t) + b$, and setting $w(t) = h(t) - b \approx an(t)$ (see ND book, section 4.2.2 for details). More generally, observations (1) and (2) suggest that lumping together (i) all the depolarizing

current conductances ($m^3(t)$) into a voltage-dependent variable $u(t) \to u_\infty(V)$ and (ii) all the hyperpolarizing current conductances ($n^4(t), h(t)$) into one abstract dynamical variable $w(t)$, with accordingly different reversal potentials $E_w, E_u$ and maximal conductances $\bar{\rho}_w, \bar{\rho}_u$, should conserve the main dynamical properties of the full system. Doing this gives us the simplified 2-dimensional *Morris-Lecar model* of action potential dynamics:

$$\tau \frac{\mathrm{d}V}{\mathrm{d}t} = -(V - E_L) - \bar{\rho}_w w(t)(V - E_w) - \bar{\rho}_u u_\infty(V)(V - E_u) + V_e(t)$$
$$\tau_w(V) \frac{\mathrm{d}w}{\mathrm{d}t} = w_\infty(V) - w$$

with $\tau_w(V) \sim \tau_h(V), \tau_n(V)$. Although this model abstracts away some of the details of the original system, it indeed seems to retain the same qualitative dynamical behaviors, evident in the similarity of its nullclines to a more precise approximation of the Hodgkin-Huxley model (fig. 2). Crucially, it is two-dimensional, so we can easily analyze it by examining its nullclines and fixed points in the $V - w$ plane.

The crucial property of this system is that its the $V$-nullcline is approximately cubic in $V$ while the $w$-nullcline is more linear[3], producing three intersections corresponding to three different fixed points when $V_e(t) = 0$ (fig. 2). We won't derive it here, but it turns out the leftmost (i.e. smallest $V$) fixed point is always stable, corresponding to the resting membrane potential. The rightmost fixed point is usually unstable (further discussion below), and the fixed point separating them is a saddle. Changing $V_e(t)$ (i.e. changing the input current $I_e(t)$) now shifts the $V$-nullcline up and down in the phase plane. Observe that shifting the $V$-nullcline up results in the stable and saddle fixed points getting closer and closer together, until eventually they merge and then disappear, leaving only the unstable fixed point at large $V$. Since the derivatives around this point are still pointing towards it, the Poincaré-Bendixson theorem[4] tells us that it must then be a limit cycle. In other words, if you increase a constant input current $I_e$ above some threshold $I_\theta$, the neuron starts spiking repeatedly. In neuroscience, $I_\theta$ is called the *rheobase* (i.e. the constant current amplitude necessary to produce spiking), and the change in number of fixed points that occurs mathematically at $I_e = I_\theta$ is called a *bifurcation*, thus making the constant input current $I$ a *bifurcation parameter*.

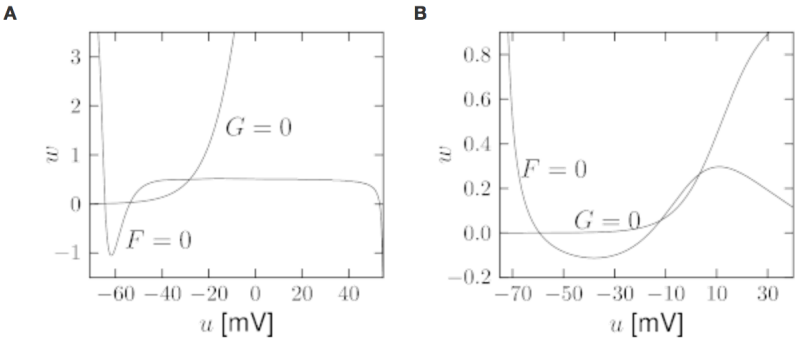



Figure 2: Copied from ND textbook fig. 4.3. **A** shows the nullclines of the Hodgkin-Huxley model rigorously reduced by setting $m(t) = m_\infty(V(t))$ and fitting a linear function $w(t) = c_h - h(t) = c_n n(t)$ such that $c_h - h(t) \approx c_n n(t)$. **B** shows the nullclines of the simplifeid Morris-Lecar model. As above, $u = V$ on abscissa.

We might then ask what the frequency of the resulting limit cycle oscillations are, to get an idea of the neuron's firing rate response to a given constant input $I$ - its so-called *gain function*. We

[3] This observation is at the heart of the classic Fitzhugh-Nagumo model, where the membrane potential dynamics are exactly a cubic function of $V$ and the dynamics of $w$ are linear in $V$:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{1}{3}z^3 + z - w$$
$$\tau_w \frac{\mathrm{d}w}{\mathrm{d}t} = b_0 + b_1 z - w$$

where I have replaced $V$ with an abstract variable $z$, reflecting the fact that these equations have completely abstracted away from the underlying physical basis of membrane potentials and de-/hyper-polarizing currents. Despite this, the reduced model shares the same qualitative features as the full one (e.g. type I, II dynamics, etc.).

[4] If (i) the fixed point is unstable and (ii) we can construct a bounding surface around it such that all derivatives on the boundary point toward its interior, then there must exist a stable limit cycle around the fixed point.

consider first the case where the rightmost fixed point is an unstable node. In this case, trajectories starting to the right of the saddle wrap around the unstable node in a counter-clockwise direction, eventually returning to the stable fixed point (fig. 3, left). When $I > I_\theta$ and the dynamics bifurcate, this behavior is conserved in the resulting limit cycle, such that the limit cycle oscillations pass through the area where the stable fixed point used to be. When $I$ is only slightly larger than $I_\theta$, the $V$-nullcline is still very near this area, such that the membrane potential derivative in that local region is very small. As a result, the oscillatory trajectories will slow down when they pass through there, reducing the spiking frequency. As $I$ gets larger and larger relative to the rheobase $I_\theta$, the slowing down of the membrane potential dynamics in this region is alleviated and the spiking frequency accordingly increases. Neuron models with such behavior are termed *type I*, characterized by a continuous monotonically increasing gain function starting from 0 at $I = I_\theta$. Intuitively, such dynamics are useful for encoding a continuous quantity, such as the overall strength of pre-synaptic input.

When the rightmost fixed point is a limit cycle to begin with, however, we get different behavior. In this case, the rightmost fixed point is a limit cycle with oscillatory trajectories that pass by just to the right of the saddle (fig. 3, right). Thus, when $I$ is increased above the rheobase and the stable and saddle nodes disappear, trajectories are pushed onto this limit cycle, which now completely avoids the area where the stable point used to be. Thus, there is no slowing down in the oscillatory trajectories and the spiking frequency immediately jumps to a high value as soon as $I > I_\theta$. Such neuron models are called *type II*, characterized by a discontinuity in their gain function, which is 0 up until $I = I_\theta$ at which point it instantly jumps up to some initial frequency substantially above 0. This behavior is useful for encoding a binary variable, endowing the neuron with switch-like (ON/OFF) dynamics.

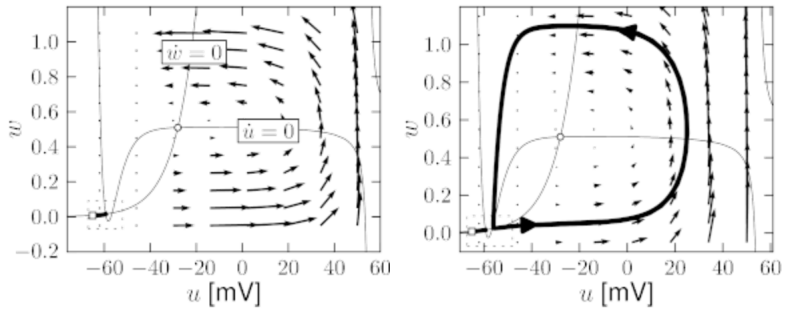



Figure 3: Copied from ND textbook figs. 4.14, 4.15. Left plot shows the derivative field with nullclines and a trajectory in the phase plane of a type I model neuron, right plot shows the same for a type II model neuron. As above, $u = V$ on abscissa.

## 2.2 Dendrites and Axons

In multi-compartmental models of neurons, we model the axons and dendrites of a neuron as *cables*. Since they are long and narrow, we assume uniformity in the radial dimension and model variations in membrane potential along the axial/longitudinal dimension:

$$V(x, t)$$

where $x$ is the axial/longitudinal position along the cable.

As we did in the single compartment case, we want to derive the temporal dynamics of the membrane potential at a given position $x$, given by:

$$C\frac{\partial V}{\partial t} = -I + I_{ext}(x, t)$$

where, as above $I_{ext}$ is an external injected current. We can think about the total current $I$ by considering a small segment of the dendrite with width $\Delta x$ centered at the longitudinal position $x$. Here, we have three sources of current:

- incoming axial current from the previous segment centered at $x - \Delta x$, given by the current at the border between the two segments: $I(x - \frac{\Delta x}{2})$

- outgoing axial current to the next segment centered at $x + \Delta x$, given by the current at that border: $I(x + \frac{\Delta x}{2})$
- membrane current generated by passive and active conductances via membrane ion channels: $I_m$

Using Ohm's law to convert current to voltage potential $I = \frac{\Delta V}{R}$, we have:

$$\begin{aligned} C\frac{\partial V}{\partial t} &= I(x - \Delta x) - I(x + \Delta x) - I_m(x) + I_{ext}(x, t) \\ &= \frac{V(x - \Delta x) - V(x)}{R_L} - \frac{V(x) - V(x + \Delta x)}{R_L} - I_m(x) + I_{ext}(x, t) \end{aligned}$$

where we now consider the intracellular *axial resistance* acting on the axial current

$$R_L = \frac{r_L l}{A} = \frac{r_L \Delta x}{\pi a^2}$$

where $l$ is the cable length (in this case equal to $\Delta x$), $a$ is the radius of the cross-section of the cable ($A = \pi a^2$ is thus the cross-sectional area), and the constant $r_L \sim 10^3 \Omega$mm is an inherent property of the neurite cytoplasm. We now approximate $V(x \pm \Delta x)$ with a second-order Taylor expansion in space:

$$\begin{aligned} C\frac{\partial V}{\partial t} &\approx \frac{\left(\left(V(x) - \Delta x \partial_x V(x) + \frac{\Delta x^2}{2}\partial_x^2 V(x)\right) - V(x)\right) - \left(V(x) - \left(V(x) + \Delta x \partial_x V(x) + \frac{\Delta x^2}{2}\partial_x^2 V(x)\right)\right)}{R_L} - I_m(x) + I_{ex} \\ &= \frac{\Delta x^2}{R_L}\frac{\partial^2 V}{\partial x^2} - I_m(x) + I_{ext}(x, t) \end{aligned}$$

Assuming the axial capacitance negligible, the capacitance term on the left-hand side becomes the membrane capacitance $C_m = c_m A = c_m 2\pi a \Delta x$ (where $A = \textit{membrane}$ area = dendrite circumference $\times$ length), so we can divide both sides by the membrane area and multiply by the specific membrane resistance to get our dynamics in terms of our good old membrane time constant:

$$\begin{aligned} \tau_m \frac{\partial V}{\partial t} &= \frac{r_m}{2\pi a \Delta x R_L}\Delta x^2 \frac{\partial^2 V}{\partial x^2} - \frac{r_m}{2\pi a \Delta x} I_m(x) + \frac{r_m}{2\pi a \Delta x} I_{ext}(x, t) \\ &= \frac{r_m a}{2 r_L}\frac{\partial^2 V}{\partial x^2} - r_m i_m(x) + r_m i_{ext}(x, t) \\ &= \lambda^2 \frac{\partial^2 V}{\partial x^2} - r_m i_m(x) + r_m i_{ext}(x, t) \end{aligned}$$

where the membrane and external currents are now in units of current per unit membrane surface area (i.e. area of surrounding cell membrane). The constant

$$\lambda = \sqrt{\frac{r_m a}{2 r_L}}$$

(in mm$^2$) is called the *electrotonic length* which, as we will see below, sets the scale of spatial (i.e. longitudinal) variation in membrane potential along the given neurite. As before, $\tau_m$ sets the scale of temporal variation.

To be able to perform some analysis on this model, we ignore the non-linear action potential-generative active conductances contained in $i_m$, leaving only the leak current $i_m = (V - E_L)/r_m$ and thus giving us the *passive cable equation*:

$$\tau_m \frac{\partial V}{\partial t} = \lambda^2 \frac{\partial^2 V}{\partial x^2} - (V - E_L) + r_m i_{ext}(x, t)$$

This simplification linearizes the dynamics, thus making it amenable to analysis. Furthermore, it is not a bad approximation whenever the membrane potential is near the resting potential or the dendrites don't have any active channels. Note that we have entirely ignored synaptic conductances until now, so our analysis is restricted to the case of external (i.e. electrode) current injection (although D&A pg. 207, top, claim that current injection can mimic the effects of a synaptic conductance).

In solving the passive cable equation, it is necessary to assume boundary conditions at branching points and end points of the cable, where the dynamics will change. Different assumptions can be made here, and in the below we take the simplest scenario: an infinite cable. Our only constraint is then that the membrane potential remain bounded for all $x, t$. While no dendrite is infinite, this is still a good approximation for sites far away from branch- or end- points of the dendrite.

We begin by considering the case of constant current injection isolated in space, i.e.

$$i_{ext}(x,t) = i_{ext}\delta(x)$$

where $x = 0$ is the exact point at which the current is injected. This will push the membrane potential to an equilibrium state given by

$$0 = \lambda^2 \frac{\partial^2 V}{\partial x^2} - (V - E_L) + r_m i_{ext}\delta(x)$$

Letting $u(x,t) = V(x,t) - E_L$, we can solve the homogenous second-order ODE (see section 7.4.2) for $x \neq 0$, where $\delta(x) = 0$:

$$\begin{aligned} &\lambda^2 \frac{\partial^2 u}{\partial x^2} - u = 0 \\ \Leftrightarrow\ & u(x) = c_1 e^{\frac{x}{\lambda}} + c_2 e^{-\frac{x}{\lambda}}, \quad x \neq 0 \end{aligned}$$

Since $u(x)$ has to be bounded for $x \to \infty, -\infty$, we then have:

$$u(x) = \begin{cases} c_1 e^{-\frac{x}{\lambda}} & \text{if } x > 0 \\ c_2 e^{\frac{x}{\lambda}} & \text{if } x < 0 \end{cases} = \Theta(x) c_1 e^{-\frac{x}{\lambda}} + \Theta(-x) c_2 e^{\frac{x}{\lambda}}$$

where $\Theta(x)$ is the Heaviside function. We therefore have a discontinuity at $x = 0$, at which point we still don't know what the membrane potential is. To find this out, we solve for $c_1, c_2$ by computing the second derivative and plugging back into the original differential equation:

$$\begin{aligned} \lambda \frac{\partial u}{\partial x} &= -\Theta(x) c_1 e^{-\frac{x}{\lambda}} + c_1\delta(x) + \Theta(-x) c_2 e^{\frac{x}{\lambda}} - c_2\delta(x) \\ &= -\Theta(x) c_1 e^{-\frac{x}{\lambda}} + \Theta(-x) c_2 e^{\frac{x}{\lambda}} + \lambda(c_1 - c_2)\delta(x) \\ \lambda^2 \frac{\partial^2 u}{\partial x^2} &= \Theta(x) c_1 e^{-\frac{x}{\lambda}} + \Theta(-x) c_2 e^{\frac{x}{\lambda}} + \lambda(-c_1 - c_2)\delta(x) + \lambda^2(c_1 - c_2)\delta'(x) \\ &= u(x) - \lambda(c_1 + c_2)\delta(x) + \lambda^2(c_1 - c_2)\delta'(x) \end{aligned}$$

The $\delta$-functions appear from the taking the derivative of the Heaviside functions. Plugging this back into the original differential equation at temporal equilibrium, we get

$$\begin{aligned} \lambda^2 \frac{\partial_x^2 u}{\partial x^2} &= u - r_m i_{ext}\delta(x) \\ \Leftrightarrow\ u - \lambda(c_1 + c_2)\delta(x) + \lambda^2(c_1 - c_2)\delta'(x) &= u - r_m i_{ext}\delta(x) \\ \Leftrightarrow\ -\lambda(c_1 + c_2)\delta(x) + \lambda^2(c_1 - c_2)\delta'(x) &= -r_m i_{ext}\delta(x) \end{aligned}$$

Since there is no term on the RHS with $\delta'(x)$, we conclude that $c_1 = c_2 = c$, giving us

$$\begin{aligned} -2c = -\frac{r_m}{\lambda} i_{ext} = -\frac{r_m}{\lambda 2\pi a} I_{ext} &= -R_\lambda I_{ext} \\ \Leftrightarrow\ c &= \frac{R_\lambda}{2} I_{ext} \\ \Rightarrow\ u(x) &= \frac{R_\lambda}{2} I_{ext} e^{-\frac{|x|}{\lambda}} \end{aligned}$$

at equilibrium (alternatively, we could have just assumed $c_1 = c_2$ on the grounds that the spatial gradient of the membrane potential should be continuous). The ratio of equilibrium potential at the injection site ($x = 0$) to the injected current $I_{ext}$ is called the *input resistance* $R_\lambda$ of the cable. This depends on a combination of the axial resistance and membrane resistance (in $R_\lambda \propto r_m, \sqrt{r_L}$). We have thus found that, when a constant current is injected into a dendrite at an infinitely small

point $x = x_0$, at equilibrium the membrane potential will drop off exponentially to each side of $x_0$, with characteristic length scale given by the electrotonic length $\lambda$:

$$V(x) - E_L = \frac{R_\lambda}{2} I_{ext} e^{-\frac{|x-x_0|}{\lambda}}$$

(see D&A fig. 6.7A for a picture). Thus, to be able to propagate a signal all the way down to the soma, dendrites can't be much longer than $\lambda$ or the current won't make it far enough before decaying to 0. This provides some insight into why real dendrites are relatively short (or have active ion channels).

Although the scenario of current being injected into an infinitely small point on the dendrite is completely unrealistic, the above analysis is still useful for understanding the membrane potential dynamics in a dendrite near resting potential $E_L$ (recall we're ignoring all active conductances). Furthermore, the solution to the passive cable equation with $i_{ext}(x) = \delta(x)$ is in fact the Green's function (section 7.4.3) for solving for the steady state of the more general case with an external current that varies smoothly over space and time. In other words, since our equation is linear, we can obtain the solution for a spatially smooth current injection by summing together the solutions to spatially isolated currents. Let $L$ be the linear operator corresponding to our passive cable equation at equilibrium, i.e.

$$Ly = \lambda^2 \frac{\partial^2 y}{\partial x^2} - y$$

Letting $u_\delta(x)$ designate the solution found above, we then have that, at equilibrium,

$$Lu_\delta(x) = -r_m i_{ext} \delta(x)$$

Consider now the case of a constant current injection varying smoothly over space:

$$\tau_m \frac{\partial u}{\partial t} = \lambda^2 \frac{\partial^2 u}{\partial x^2} - u + r_m f(x)$$

(where $f(x)$ is in units of current per unit area). To solve for distribution of membrane potential over space at the temporal equilbrium, we set $\frac{\partial u}{\partial t} = 0$ and use the *Green's function* (section 7.4.3) given by $u_\delta$:

$$\begin{aligned}
\lambda^2 \frac{\partial_x^2 u}{\partial x^2} - u &= -r_m f(x) \\
\Leftrightarrow Lu &= -r_m f(x) \\
&= \int_{-\infty}^{\infty} -\delta(x - x') r_m f(x') \mathrm{d}x' \\
&= \int_{-\infty}^{\infty} \frac{L u_\delta(x - x')}{i_{ext}} f(x') \mathrm{d}x' \\
&= L \int_{-\infty}^{\infty} \frac{u_\delta(x - x')}{i_{ext}} f(x') \mathrm{d}x' \\
\Leftarrow u(x) &= \frac{1}{i_{ext}} \int_{-\infty}^{\infty} u_\delta(x - x') f(x') \mathrm{d}x'
\end{aligned}$$

where we were able to go from the fourth to the fifth line since integration and differentiation are both linear operators, and $L$ consisted of differentiating with respect to $x$ whereas the integral was over a different variable $x'$. The awkward left arrow on the last line makes the rather technical point that we have not shown that this is *the* unique solution to the ODE, only that it is *a* solution (I believe a boundary condition at $x = 0$ would suffice to get a unique solution). Recalling that $u_\delta(x) \propto e^{-\frac{|x|}{\lambda}}$ is just a rising and then decaying exponential, the resulting spatial distribution at equilibrium will simply look like a smoothed $f(x)$ as a result of the convolution with $u_\delta(x)$.

The next case to consider is a pulse of injected current isolated in space and time:

$$\tau_m \frac{\partial u}{\partial t} = \lambda^2 \frac{\partial^2 u}{\partial x^2} - u + r_m i_{ext} \delta(x) \delta(t)$$

We solve this by first taking the Fourier transform in space, which gives us a simple first-order ODE:

$$\begin{aligned}\tau_m \frac{\partial}{\partial t} U(\omega) &= -\lambda^2\omega^2 U(\omega) - U(\omega) + r_m i_{ext}\delta(t) \\ \Leftrightarrow \frac{\partial}{\partial t} U(\omega) + \frac{\lambda^2\omega^2+1}{\tau_m} U(\omega) &= \frac{r_m i_{ext}}{\tau_m}\delta(t)\end{aligned}$$

where we used the fact that the Fourier transform of a derivative $\frac{d^n}{dx^n}f(x)$ is equal to $(\omega i)^n F(\omega)$. Solving this, we get:

$$\begin{aligned}U(\omega,t) &= U(\omega,0)e^{-\frac{\lambda^2\omega^2+1}{\tau_m}t} + \frac{r_m i_{ext}}{\tau_m}\Theta(t)e^{-\frac{\lambda^2\omega^2+1}{\tau_m}t} \\ &\approx \frac{r_m i_{ext}}{\tau_m}\Theta(t)e^{-\frac{\lambda^2\omega^2+1}{\tau_m}t}\end{aligned}$$

by setting $u(x,0) = 0 \Rightarrow U(\omega,0) = 0$. Taking the inverse Fourier transform of both sides to return to the spatial domain, we note that the exponential term on the RHS is squared exponential in $\omega$, meaning we can easily compute its inverse Fourier transform by putting it into Gaussian form:

$$\begin{aligned}U(\omega,t) &= \frac{r_m i_{ext}}{\tau_m}\Theta(t)e^{-\frac{t}{\tau_m}}e^{-\frac{\lambda^2 t}{\tau_m}\omega^2} \\ &\overset{\omega\to 2\pi k}{=} \frac{r_m i_{ext}}{\tau_m}\Theta(t)e^{-\frac{t}{\tau_m}}\sqrt{\frac{\tau_m}{4\pi\lambda^2 t}}\sqrt{\frac{4\pi\lambda^2 t}{\tau_m}}e^{-\frac{4\lambda^2 t}{\tau_m}\pi^2 k^2} \\ &= \frac{r_m i_{ext}}{\tau_m\sqrt{\pi B t}}\Theta(t)e^{-\frac{t}{\tau_m}}\sqrt{\pi B t}e^{-Bt\pi^2 k^2} \\ \Rightarrow u(x,t) &= \frac{r_m i_{ext}}{\tau_m\sqrt{\pi B t}}\Theta(t)e^{-\frac{t}{\tau_m}}e^{-\frac{x^2}{Bt}} \\ &= \frac{R_\lambda I_{ext}}{\sqrt{4\pi\tau_m t}}\Theta(t)e^{-\frac{t}{\tau_m}}e^{-\frac{\tau_m}{4\lambda^2 t}x^2}\end{aligned}$$

where

$$B = \frac{4\lambda^2}{\tau_m}$$

and we took the change of variables $\omega \to 2\pi k$, such that $k$ is in units of frequency (i.e. inverse units of $x$) and we could exploit our formula for the Fourier transform of a Gaussian function (section 7.6). We thus see that the pulse of current injection decays with distance from the injection site as a Gaussian with width $\sqrt{Bt} \propto \lambda$, which expands over time as $\sqrt{t/\tau_m}$ and the peak decays as $e^{-\frac{t}{\tau_m}}/\sqrt{t}$. In other words, the dynamics of the current injection are a spreading Gaussian in space with integral decaying exponentially in time (see D&A fig 6.7B for a picture). This solution in turn provides the Green's function to solve for the propagation of a current injection that varies over space and time.

If we look at a site $x$ away from the current injection site, the current as a function of time looks like a difference of exponentials (D&A fig 6.8A) with the peak at some $t^* > 0$, later for sites further away. We can thus try to compute a kind of "velocity" of current propagation by computing the amount of time it will take for the current at site $x$ to peak, and dividing the distance $x$ from the current injection site (i.e. the distance travelled) by the time to peak. The time of the peak $t^*$ is easily computed by setting the time derivative of the logarithm of $u(x,t)$ to 0 (assuming $t > 0$):

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t^*}\left[-\frac{t}{\tau_m} - \frac{\tau_m x^2}{4\lambda^2 t} - \frac{1}{2}\log t + \text{const. w.r.t. } t\right] \\ &= -\frac{1}{\tau_m} + \frac{\tau_m x^2}{4\lambda^2 t^{*2}} - \frac{1}{2t^*} \\ &= \frac{4\lambda^2}{\tau_m}t^{*2} + 2\lambda^2 t^* - \tau_m x^2 \\ \Leftrightarrow t^* &= \frac{-\lambda \pm \sqrt{\lambda^2 + 4x^2}}{4\lambda}\tau_m \\ &= \frac{\tau_m}{4}\left(\sqrt{1 + 4x^2/\lambda^2} - 1\right)\end{aligned}$$

(ignoring the negative root of the quadratic, since $t^* > 0$) which, in the limit of large $x$ is:

$$t^* \approx \frac{\tau_m}{4}\left(\frac{2x}{\lambda}\right) = \frac{\tau_m x}{2\lambda}$$

In this limit, the velocity of current propagation is then:

$$v_{dendrite} = \frac{x}{t^*} \approx \frac{2\lambda}{\tau_m}$$

where the approximation is good for sites far away from the injection site, e.g. the soma when current is injected at a distal dendrite.

Axons, however, often need to propagate signals over long distances and therefore require higher speeds of propagation. Given that $r_L$ and $c_m$ are intrinsic properties of the cell cytoplasm and phospholipid bilayer, the two parameters we can manipulate to achieve higher speeds are $a$ (axon radius) and $r_m$ (membrane resistance). It turns out the mammalian brain does both. To change $r_m$, long-range projecting axons are often *myelinated*: they are wrapped with layers of cell membrane (*myelin*) that effectively increase the membrane resistance. We model this by taking $r_m \to \infty$. Rearranging the passive cable equation to take this limit and then using the same strategy as above to solve for the propagation of a pulse of injected current (Fourier transform in space $\to$ solve differential equation in time $\to$ inverse Fourier transform of a Gaussian), we get:

$$\begin{aligned}
c_m\frac{\partial V}{\partial t} &= \frac{\lambda^2}{r_m}\frac{\partial^2 V}{\partial x^2} - \frac{V - E_L}{r_m} + i_{ext}\delta(x)\delta(t) \\
&= \frac{a}{2r_L}\frac{\partial^2 V}{\partial x^2} - \frac{V - E_L}{r_m} + i_{ext}\delta(x)\delta(t) \\
\Rightarrow \lim_{r_m\to\infty}\frac{\partial V}{\partial t} &= \frac{a}{2r_L c_m}\frac{\partial^2 V}{\partial x^2} + i_{ext}\delta(x)\delta(t) \\
\Rightarrow V(x,t) &= \frac{i_{ext}}{\sqrt{\pi D t}}\Theta(t)e^{-\frac{x^2}{Dt}} \\
D &= \frac{2a}{r_L c_m}
\end{aligned}$$

Note the lack of a term decaying exponentially with time, meaning that in this setting the signal propagates as a Gaussian spreading in time, with constant integral (an intuitive result from the fact that myelination effectively eliminates the leak current). This slowing down of the signal decay results in faster "velocity" of the propagating signal in the axon, which we can compute as above:

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t^*}\left[-\frac{r_L c_m x^2}{2at} - \frac{1}{2}\log t + \text{const. w.r.t. } t\right] \\
&= \frac{r_L c_m x^2}{2at^{*2}} - \frac{1}{2t^*} \\
\Leftrightarrow t^* &= \frac{r_L c_m x^2}{a} \\
\Rightarrow v_{axon} &= \frac{a}{r_L c_m x}
\end{aligned}$$

This looks like bad news: $v_{axon} \propto 1/x$ so long axons will have very slow signal propagation to their terminals. To deal with this, it turns out that in mammalian neural systems $a \propto L$. This means that for long and (therefore) thick myelinated axons,

PEL said this in lecture - reference??

$$v_{axon} = \frac{1}{r_L c_m} = \frac{2\pi a}{r_L C_m}$$

Thus, we have that (approximately)

$$\begin{aligned}
v_{dendrite} &\propto \sqrt{a} \\
v_{axon} &\propto a
\end{aligned}$$

For further discussion of such wiring principles at play in the mammalian brain, see Chklovskii et al., 2002.

Note, however, that the spatial decay of the signal remains the same in axons as in dendrites, since the Gaussians have the same width:

$$B = \frac{4\lambda^2}{\tau_m} = \frac{4r_m a}{2r_L r_m c_m} = \frac{2a}{r_L c_m} = D$$

So, although a signal originating from the soma may propagate faster down an axon, it will still decay to 0 for any distances much further than about $2\sqrt{Dt}$. Since axons need to be long to project to different brain areas, they deal with this problem by separating segments of myelination with so-called *nodes of Ranvier* where there is a high concentration of active $Na^+$ channels that can initiate an action potential if the membrane potential gets high enough. This is called *saltatory conductance*, since the action potential "jumps" (*salta*, in Spanish) from one node to the next.

## 2.3 Synaptic Transmission

Synaptic transmission is a three-stage process:

1. An action potential arrives at the pre-synaptic terminal, thus opening voltage-dependent calcium ion ($Ca^{2+}$) channels and leading to an increase in the intracellular $Ca^{2+}$ concentration.
2. With probability $P_{rel}$, this triggers the fusion of vesicles containing neurotransmitter to the cell membrane, leading to the release of the neurotransmitter into the synaptic cleft.
3. With probability $p_j$, neurotransmitter binds to type-$j$ receptors on the post-synaptic cell membrane, causing type-$j$ ion channels to open. If the resulting post-synaptic membrane current is strong enough (summing over all $j$ at the synapse), the membrane potential may rise above threshold and trigger an action potential in the post-synaptic cell.

Crucially, steps 2 and 3 are stochastic, so our synaptic conductance-based model for the post-synaptic membrane current at synapse $s$ is

$$i_s = \xi_s \sum_j \bar{g}_j p_j (V - E_j)$$

$$\xi_s = \begin{cases} 1 & \text{with probability } P_{rel} \\ 0 & \text{with probability } 1 - P_{rel} \end{cases}$$

where $j$ indexes different neurotransmitter-dependent ion channel types on the post-synaptic membrane at the given synaptic cleft. Our notation follows Hodgkin-Huxley model conventions, with $\bar{g}_j$ (= conductance of open channel $j$ × density of channel $j$) and $E_j$ (= reversal potential of ions channel $j$ is permeable to) as constants and $p_j(t)$ (= probability of a $j$-channel being open) modelled as a gating variable with (two-state Markov model) dynamics

$$\frac{\mathrm{d}p_j}{\mathrm{d}t} = \alpha_j(C_j)(1 - p_j) - \beta_j(C_j)p_j$$

$$\Leftrightarrow \tau_j(C_j)\frac{\mathrm{d}p_j}{\mathrm{d}t} = p_\infty^{(j)}(C_j) - p_j$$

$$\tau_j(C_j) = \frac{1}{\alpha_j(C_j) + \beta_j(C_j)}, \quad p_\infty^{(j)}(C_j) = \frac{\alpha_j(C_j)}{\alpha_j(C_j) + \beta_j(C_j)}$$

where $C_j$ is the concentration (in the synaptic cleft) of the neurotransmitter that activates channel $j$ and $E_j$ designates the reversal potential for the ions that channel $j$ is permeable to. $\alpha_j$ and $\beta_j$ respectively refer to the rate of binding and unbinding of neurotransmitter to the given receptor type. Typically, at a given synapse there will only be one type of neurotransmitter being released by the pre-synaptic cell (the strongest version of *Dale's Law*: every neuron releases only one type of neurotransmitter at all its synaptic terminals), but I will continue with the general case.

Solving the equation above gives us

$$p_j(t) = p_\infty^{(j)}(C_j) + (p_j(0) - p_\infty^{(j)}(C_j))e^{-\frac{t}{\tau_j(C_j)}}$$

A useful simplification here is to assume $\beta_j$ to be a small constant, and to set $\alpha_j(C_j) \propto C_j^k$ with some exponent $k$ such that neurotransmitter binding rate is highly dependent on the concentration

of neurotransmitter in the synaptic cleft, with $\alpha_j(0) = 0$. We then model the concentration of neurotransmitter $C_j(t)$ as a square wave

$$C_j(t) = \bar{C}_j \Theta(t)\Theta(T-t)$$

with large $\bar{C}_j$ so that $p_\infty^{(j)}(C_j) \approx 1$ at times $t \in [0, T]$. In reality, after neurotransmitter is released into the synaptic cleft, it is quickly removed via enzyme-mediated degradation as well as through diffusion, making the square wave a reasonable approximation. This results in the following solution:

$$p_j(t) = \begin{cases} 1 - (1 - p_j(0))e^{-(\alpha_j(\bar{C}_j)+\beta_j)t} & \text{if } 0 \le t \le T \\ p_j(T)e^{-\beta_j t} & \text{if } t > T \end{cases}$$

which consists of a saturating (to 1) rising exponential with time constant $\tau_{rise} = \frac{1}{\alpha_j(\bar{C}_j)+\beta_j}$ at times $t \in [0, T]$ followed by an exponential decay with time constant $\tau_{decay} = \frac{1}{\beta_j}$, where time $t = 0$ indicates the moment at which neurotransmitter is released into synaptic cleft.

A common further simplification is to assume instantaneous neurotransmitter release and removal by letting $T \to 0$ so that $C_j(t) \to \bar{C}_j \delta(t)$, and setting

$$p_j(0^+) = p_j(T) = p_j(0^-) + (1 - p_j(0^-))p_j^{max}$$

where $p_j(0^+), p_j(0^-)$ are the probability of channel $j$ being open at the exact moment of and just prior to neurotransmitter release, respectively. Here, we've set $p_j(0^+)$ to its maximum in the previous more realistic model, given by $p_j(T)$. In this case, $p_j^{max} = (1 - e^{-(\alpha_j(\bar{C}_j)+\beta_j)T})$. Generalizing this model to arbitrary pre-synaptic spike times $\{t_k\}$ gives us the following synaptic conductance dynamics[5]:

$$\frac{\mathrm{d}p_j}{\mathrm{d}t} = -\beta_j p_j + (1 - p_j)p_j^{max} \sum_k \xi_k \delta(t - t_k)$$

$$\xi_k = \begin{cases} 1 & \text{with probability } P_{rel} \\ 0 & \text{with probability } 1 - P_{rel} \end{cases}$$

Note that $\xi_k, t_k$ are not indexed by $j$ (in fact they should be indexed by the particular synapse $s$) - all receptors $j$ at this synapse share the same pre-synaptic spike times and neurotransmitter release probability. In this case, we drop the $\xi_s$ term in our equation for the post-synaptic membrane current $i_s$, since the stochastic vesicle release component is now implicit in the channel opening probabilities.

A more phenomenological model of the synaptic conductance is the difference-of-exponentials

$$p_j(t) = p_j^{max} B \left( e^{-\frac{t}{\tau_1}} - e^{-\frac{t}{\tau_2}} \right)$$

$$B = \left( \left( \frac{\tau_2}{\tau_1} \right)^{\frac{\tau_{rise}}{\tau_1}} - \left( \frac{\tau_2}{\tau_1} \right)^{\frac{\tau_{rise}}{\tau_2}} \right)^{-1}, \quad \tau_1 > \tau_2$$

with rise time $\tau_{rise} = \frac{\tau_1 \tau_2}{\tau_1 - \tau_2}$ and decay time $\tau_{decay} = \tau_1$. The normalizer $B$ simply enforces that the peak of the conductance be $p_j^{max}$, which occurs at $t^* = \tau_{rise} \log \frac{\tau_1}{\tau_2}$. Another more simplified phenomenlogical model is the $\alpha$-function

$$p_j(t) = \frac{p_j^{max} t}{\tau_j} e^{1 - t/\tau_j}$$

which reaches its maximum $p_j^{max}$ at $t^* = \tau_j$, with decay time $\tau_{decay} = \tau_j$. These two models are useful for neurotransmitter receptors with slower rise times (e.g. $\mathrm{GABA}_B$, NMDA).

Note that under certain simplifications, we now have a full model of brain activity. Specifically, if we assume only one type of post-synaptic channel at each synapse and ignore the dynamics of

[5] As hinted just above, this simple looking differential equation hides an analytical obstacle in that it requires evaluating $p_j(t_k)$, which is impossible to evaluate since $p_j$ is in the middle of an infinite slope jump at this point because of the contribution of the $\delta$-function $\delta(t - t_k)$. My derivation implies circumventing this problem by simply evaluating $p_j(t_k)$ at time $t_k^-$, just before the jump. This is called an Itô integral. It is worth noting, however, that an alternative approach would be evaluating $p_j(t_k)$ in the middle of the jump, called a Stratonovich integral.

dendrites and axons, the following equation gives us the membrane potential dynamics of any given neuron $i$:

$$\tau_m \frac{\mathrm{d}V_i}{\mathrm{d}t} = -(V_i - E_L) - [\text{HH active currents}] - \sum_j \bar{g}_{ij}\xi_{ij}p_{ij}(V_i - E_j)$$

$$\tau_{ij}(C_{ij})\frac{\mathrm{d}p_{ij}}{\mathrm{d}t} = p_\infty^{(ij)}(C_{ij}) - p_{ij} \quad [\text{insert favorite synaptic conductance model here}]$$

$$\xi_{ij} = \begin{cases} 1 & \text{with probability } P_{rel}^{(ij)} \\ 0 & \text{with probability } 1 - P_{rel}^{(ij)} \end{cases}$$

where $j$ indexes all synapses $ij$ onto neuron $i$. This is a conductance-based model of the brain, since we are explicitly modelling the conductances in $\bar{g}_{ij}\xi_{ij}p_{ij}$. Alternatively (as we will do below in section 3), we could simplify this to a current-based model by absorbing $\bar{g}_{ij}\xi_{ij}(V_i - E_j)$ into one term $W_{ij}$ such that the synaptic inputs to model $i$ (i.e. the last term in the equation for $\frac{\mathrm{d}V_i}{\mathrm{d}t}$) are modelled as input currents.

In fact, this is often not such a bad simplification since we can classify all neurotransmitters as either excitatory or inhibitory, depending on the polarity of the reversal potential of the ion currents generated by their respective post-synaptic receptors (i.e. greater/less than the resting potential $E_L$ for excitatory/inhibitory). A great diversity exists within each of these classes, but it turns out we can reproduce many characteristic properties of brain dynamics with just these two classes of synapse (section 3). Two such neurotransmitters ubiquitous in the neocortex are *glutamate* ($E \sim 0$mV) and *GABA* ($E \sim -100$mV), with post-synaptic receptors given in the table.

| Neurotransmitter | Receptor | Time constant | Ions |
|---|---|---|---|
| Glutamate | AMPA | fast ($\sim$ 1ms) | cations |
| | NMDA | slow | cations, including $Ca^{2+}$ |
| GABA | $GABA_A$ | fast | $Cl^-$ conductance |
| | $GABA_B$ | slow | $K^+$ conductance |

These are all *ionotropic* receptors, since the binding of a neurotransmitter to them automatically opens an ion channel attached to the receptor. Other receptors can be *metabotropic*, meaning that they trigger an intracllular signalling cascade that results in a certain type of channel opening.

### 2.3.1 Short-term Synaptic Plasticity: Synaptic Depression and Facilitation

The current outlook on synaptic plasticity is that, at a synapse $ij$, the maximal conductance term $\bar{g}_{ij}$ changes on long timescales (e.g. by increasing/decreasing density of receptors) whereas the neurotransmitter release probability $P_{rel}^{(ij)}$ change on both short and long timescales. On short timescales one of two things can happen:

- *Synaptic depression:* post-synaptic potential temporarily decreases with repeated high frequency pre-synaptic spikes, since the stock of readily available neurotransmitter in the pre-synaptic axon terminal has been depleted, thus lowering the probability of vesicle release on the next spike.
- *Synaptic facilitation:* post-synaptic potential temporarily increases with repeated high frequency pre-synaptic spikes, since this leads to a high influx of calcium $Ca^{2+}$ ions into the pre-synaptic axon terminal, thus increasing the probability of vesicle release on the next spike.

We can thus model both synaptic depression and facilitation with a two-dimensional system of ODEs explicitly modelling the dynamics of calcium ion concentration $[Ca^{2+}]$ in the pre-synaptic terminal and the number of vesicles $M$ ready for release:

$$P_{rel} = \mathrm{f}([Ca^{2+}], M)$$

$$\tau_{Ca}\frac{\mathrm{d}[Ca^{2+}]}{\mathrm{d}t} = -[Ca^{2+}] + \alpha\sum_k \delta(t - t_k)$$

$$\tau_M \frac{\mathrm{d}M}{\mathrm{d}t} = M_0 - M - \sum_k \xi_k \delta(t - t_k)$$

with some complicated f($[Ca^{2+}], M$) that is presumably monotonically increasing in $[Ca^{2+}], M$.

An alternative is to abstract away and forget the calcium and vesicle number dynamics by directly modelling the temporal dynamics of $P_{rel}$:

$$\tau_{rel}\frac{\mathrm{d}P_{rel}}{\mathrm{d}t} = P_0 - P_{rel} + \tau_{rel}\sum_k \delta(t-t_k) \times \begin{cases} -\xi_k(1-f_D)P_{rel} \\ f_F(1-P_{rel}) \end{cases}$$

where $f_F, f_D \in [0,1]$ (larger $f_F$ for stronger facilitation, smaller $f_D$ for stronger depression), $k$ indexes pre-synaptic spikes, and $\xi_k$ is our usual stochastic variable taking on 1 with probability $P_{rel}(t)$ and 0 otherwise (representing whether or not vesicles were released upon pre-synaptic spike $k$). In other words, the release probability decays exponentially to its equilibrium value $P_0$, updating itself every time a pre-synaptic action potential arrives at the axon terminal according to either a synaptic depression or facilitation update rule. Because there is a $P_{rel}$ term multiplied by $\delta$-functions on the right hand side, analyzing this equation becomes a mess (see footnote 5), so it is actually easier work with the explicit update rules:

$$\begin{aligned} \tau_{rel}\frac{\mathrm{d}P_{rel}}{\mathrm{d}t} &= P_0 - P_{rel} \\ P_{rel} &\to \xi_k f_D P_{rel} + (1-\xi_k)P_{rel} && \text{[synaptic depression]} \\ P_{rel} &\to P_{rel} + f_F(1-P_{rel}) && \text{[synaptic facilitation]} \end{aligned}$$

where the updates occur upon a pre-synaptic spike arriving at the axon terminal.

Synaptic depression can be particularly useful for normalizing synaptic inputs and for detecting changes in firing rate. We can see this by setting $\xi_k = 1$ to allow perfectly reliable vesicle fusion and neurotransmitter release on every pre-synaptic spike and then computing the steady state $\langle P_{rel}\rangle$ averaged over pre-synaptic spikes drawn from some homogenous Poisson process with rate $r$. Suppose the release probability is at this average steady state, $P_{rel} = \langle P_{rel}\rangle$, when a single pre-synaptic spike occurs at time $t_k$ so that:

$$P_{rel} \to f_D\langle P_{rel}\rangle$$

Solving our ODE, the release probability at the time of the next spike $t_{k+1}$ is then given by

$$P_{rel}(t_{k+1}) = P_0 + (f_D\langle P_{rel}\rangle - P_0)e^{-\frac{t_{k+1}-t_k}{\tau_{rel}}}$$

Having defined $\langle P_{rel}\rangle$ as the average steady state, averaging over spike times $t_{k+1}$ should give us $\langle P_{rel}(t_{k+1})\rangle = \langle P_{rel}\rangle$. In other words, the release probability should on average decay to $\langle P_{rel}\rangle$ by the time the next spike arrives (by definition). Given that the pre-synaptic spike times are drawn from a homogenous Poisson process, we can directly compute this average by averaging over the exponentially distributed inter-spike intervals:

$$\begin{aligned} \langle P_{rel}(t_{k+1})\rangle &= P_0 + (f_D\langle P_{rel}\rangle - P_0)\int_{t_k}^{\infty} P(t_{k+1}-t_k)e^{-\frac{t_{k+1}-t_k}{\tau_{rel}}}\mathrm{d}t_{k+1} \\ &= P_0 + (f_D\langle P_{rel}\rangle - P_0)\int_0^{\infty} re^{-r\tau}e^{-\frac{\tau}{\tau_{rel}}}\mathrm{d}\tau \\ &= P_0 + (f_D\langle P_{rel}\rangle - P_0)r\int_0^{\infty} e^{-\tau\left(\frac{r\tau_{rel}+1}{\tau_{rel}}\right)}\mathrm{d}\tau \\ &= P_0 + (f_D\langle P_{rel}\rangle - P_0)\frac{r\tau_{rel}}{r\tau_{rel}+1} \end{aligned}$$

Setting $\langle P_{rel}(t_{k+1})\rangle = \langle P_{rel}\rangle$, we then solve for the average steady state $\langle P_{rel}\rangle$:

$$\langle P_{rel}\rangle = \frac{P_0\left(1-\frac{r\tau_{rel}}{r\tau_{rel}+1}\right)}{1-f_D\frac{r\tau_{rel}}{r\tau_{rel}+1}} = \frac{P_0}{(1-f_D)r\tau_{rel}+1}$$

We thus see that, at high pre-synaptic firing rates $r$, the release probability scales with $\frac{1}{r}$. This means that the rate of arriving post-synaptic potentials, given by $rP_{rel}$, remains approximately constant with respect to the pre-synaptic firing rate $r$ (at steady state). Synaptic depression thus serves as a mechanism for normalizing pre-synaptic inputs (with potentially different firing rates)

on different synapses to the same synaptic transmission rate (and hence the same time-averaged post-synaptic potential amplitude, assuming individual post-synaptic potential amplitudes to be the same across synapses), at least in the regime of high pre-synaptic firing.

Of course, this also means that the synapse cannot convey any information about smooth changes in pre-synaptic firing rate, on the timescale of $\tau_{rel}$. Faster changes $r \to r + \Delta r$, however, will be detected since it takes $\mathcal{O}(\tau_{rel})$ time for $P_{rel}$ to reach its new steady state. Before reaching it, the synaptic transmission rate will thus transiently rise to

$$(r + \Delta r)\langle P_{rel}\rangle = \frac{(r + \Delta r)P_0}{(1 - f_D)r\tau_{rel} + 1}$$

which, in the limit of high pre-synaptic firing rates, is $\mathcal{O}(\frac{r+\Delta r}{r})$. In other words, the resulting increase in synaptic transmission rate is proportional to the relative, rather than absolute, increase in pre-synaptic firing rate. A synapse can therefore use synaptic depression to encode the relative magnitude of transient changes in the pre-synaptic firing rate.

### 2.3.2 NMDA-mediated plasticity

One of the few postulated mechanisms for long-term plasticity (section 4) is the unblocking of NMDA receptors via back-propagating action potentials. NMDA receptors are unique in that they have sites that magnesium $Mg^{2+}$ ions bind to, thus blocking the receptor in a voltage-dependent fashion. Namely, since $Mg^{2+}$ ions have a positive charge, a high membrane potential at the post-synaptic cell/dendrite will repel them, thus unblocking the NMDA receptors at the synapse. We can thus model NMDA receptor conductance by incorporating a scaling factor that is sigmoidal in the membrane potential $V$, this sigmoidal relationship mediated by concentration of magnesium ions in the synaptic cleft $[Mg^{2+}]$:

$$i_{\text{NMDA}} = -\frac{\bar{g}_{\text{NMDA}} p_{\text{NMDA}}}{1 + \frac{[Mg^{2+}]}{3.57} e^{-\frac{V}{16.8}}}(V - E_{\text{NMDA}})$$

(where the vesicle release probability is implicit in $p_{\text{NMDA}}$).

Importantly, it turns out that when a neuron spikes, action potentials often "back-propagate" up the dendrites, thus increasing the membrane potential near the synapses and unblocking NMDA receptors. Furthermore, NMDA receptors are permeable to $Ca^{2+}$ ions, which trigger long-term changes at the synapse by signalling the cell to (1) open more NMDA channels and (2) produce and insert new AMPA channels. This process is called NMDA-mediated plasticity. Note that NMDA-mediated plasticity at a synpase can also be triggered by increases in the membrane potential resulting from the depolarization of neighboring synapses, leading to what is called *heterosynaptic plasticity* (more on the functional implications of this in section 4).

Because it requires both pre- and post-synaptic activity, NMDA-dependent plasticity is thought to play a major role in Hebbian learning. Indeed, it has been experimentally observed that in many cases synaptic plasticity stops when NMDA channels are blocked. From a neurocomputational point of view, NMDA receptors can act as coincidence detectors, since they are most active under simultaneous post- and pre- synaptic depolarization.

# 3 Networks

## 3.1 Mean-Field Analysis of Spiking Networks

We now consider the simplified network model given by the differential equations

$$\tau_m \frac{\mathrm{d}V_i}{\mathrm{d}t} = f_i(V_i, t) - \sum_{j \neq i} m_{ij} g_j(t)(V_i - E_j) \tag{1}$$

$$\tau_s \frac{\mathrm{d}g_j}{\mathrm{d}t} = -g_j + \sum_k \delta(t - t_k^{(j)}) \tag{2}$$

where $f_i(V_i, t)$ represents the single neuron dynamics (e.g. leak current, Hodgkin-Huxley currents) and $m_{ij}$ is an abstract variable meant to represent the contributions of the neurotransmitter release probability, the ion channel density, and the open ion channel conductance (i.e. combination of

$\xi_j \sim P_{rel}$ and $\bar{g}_i$) across all synapses between pre-synaptic neuron $j$ and post-synaptic neuron $i$. The pre-synaptic neuron $j$ is assumed to release the same neurotransmitter at all its axon terminals, with associated reversal potential $E_j$ (+ve for excitatory, -ve for inhibitory). The dynamics of $g_j$ are meant to emulate the dynamics of a synaptic conductance, under the simplification that each pre-synaptic spike at time $t_k^{(j)}$ instantly triggers neurotransmitter release at all its axon terminals, leading to an instant rise in the neurotransmitter concentration at the synaptic cleft, modelled by a $\delta$-function in equation 2.

So we now have a set of $2N$ equations comprising a dynamical system that we have reason to believe could be a good description of a vanilla neural circuit of $N$ neurons in the brain. We might then ask: what kinds of dynamical behaviors can we expect from such a system? Can we derive any general principles about network dynamics from this set of equations? Unfortunately, the non-linear terms $f_i(V_i, t)$ and $g_j(t)V_i(t)$ in equation 1 make this system very hard to analyze, so we will require further simplifications to make any analytical headway.

The first such simplification is to simplify the non-linearities in equation 1. First, we get rid of the non-linear interaction $g_j(t)V_i(t)$ between the membrane potential and synaptic conductance by moving from a *conductance-based model* to a *current-based model*. Namely, we we approximate the post-synaptic membrane potential in this interaction term by its temporal mean: $g_j(t)V_i(t) \to g_j(t)\bar{V}_i$, allowing us to rewrite equation 1 as

$$\tau_m \frac{\mathrm{d}V_i}{\mathrm{d}t} = f_i(V_i, t) + \sum_{j \neq i} w_{ij} g_j(t)$$

where $w_{ij} = -m_{ij}(\bar{V}_i - E_j)$ is interpreted as an approximate "synaptic weight" from pre-synaptic neuron $j$ onto post-synaptic neuron $i$.

We now focus on the *synaptic drive*

$$h_i(t) = \sum_j w_{ij} g_j(t)$$

(henceforth all sums over pre-synaptic neurons $j \neq i$ will be written shorthand as sums over $j$), which is a function of time by virtue of the temporal dynamics of $g_j(t)$ (equation 2). Note that these dynamics are such that the total current contribution of a single pre-synaptic spike $k$ from neuron $j$ integrates to 1:

$$g_j(t) = \Theta(t - t_k^{(j)}) \frac{1}{\tau_s} e^{-\frac{-(t-t_k)}{\tau_s}}$$

$$\Rightarrow \int_0^\infty g_j(t)\mathrm{d}t = \int_{t_k^{(j)}}^\infty \frac{1}{\tau_s} e^{-\frac{-(t-t_k)}{\tau_s}} \mathrm{d}t = \left[-e^{-\frac{-(t-t_k)}{\tau_s}}\right]_{t_k^{(j)}}^\infty = 1$$

Thus, its temporal mean $\bar{g}_j = \langle g_j(t) \rangle_t$ is in fact the mean firing rate $\nu_j$ of neuron $j$:

$$\begin{aligned}
\langle g_j(t) \rangle_t &= \lim_{T \to \infty} \frac{1}{T} \int_0^T g_j(t)\mathrm{d}t \\
&= \lim_{T \to \infty} \frac{1}{T} \sum_k \int_{t_k^{(j)}}^\infty \frac{1}{\tau_s} e^{-\frac{-(t-t_k)}{\tau_s}} \mathrm{d}t \\
&= \lim_{T \to \infty} \frac{1}{T} n_j(T) \equiv \nu_j
\end{aligned}$$

where $n_j(T)$ designates the number of emitted spikes up until time $T$. We can thus rewrite the synaptic drive to neuron $i$ in terms of mean firing rates:

$$\begin{aligned}
h_i(t) &= \sum_j w_{ij} g_j(t) \\
&= \sum_j w_{ij} (\bar{g}_j + \delta g_j(t)) \\
&= \underbrace{\sum_j w_{ij} \nu_j}_{\text{"quenched noise"}} + \underbrace{\sum_j w_{ij} \delta g_j(t)}_{\text{"dynamic noise"}} \\
&= \bar{h}_i + \delta h_i(t)
\end{aligned}$$

where $\delta g_j(t) \equiv g_j(t) - \bar{g}_j$ are 0-mean fluctuations of each $g_j(t)$ around its temporal mean $\bar{g}_j$, such that

$$\langle \delta h_i(t) \rangle_t = \sum_j w_{ij} \langle \delta g_j(t) \rangle_t = 0$$

We can thus express the synaptic drive $h_i(t)$ as 0-mean temporal fluctuations $\delta h_i(t)$ (the "dynamic noise") around a temporal mean $\bar{h}_i$ (the "quenched noise" - noisy over neurons rather than over time).

This inspires the following idea: rather than analyzing actual trajectories of our high-dimensional non-linear system (which requires solving a set of $2N$ non-linear ODEs - a super hard problem), let's instead try to solve the following two possibly easier *statistical* problems:

1. characterize the distribution of *time-averaged* activity over all neurons in the network (arising from the quenched noise component of the synaptic drive)

2. characterize temporal correlations in activity (arising from the dynamic noise component of the synaptic drive)

Solving these problems naturally won't give us temporal trajectories or anything like what we'd get from solving the ODEs, but it will still prove useful to gain some insight into the different dynamical regimes such a system can exhibit.

Here, we'll focus mainly on problem #1: characterizing the distribution over time-averaged firing rates in the population $\{\nu_i\}_{i=1}^N$, where "characterizing" will simply entail computing the moments of the distribution $\langle \nu^\ell \rangle$. Our approach to doing this stems from the following two observations:

I. the time-averaged synaptic drive $\bar{h}_i = \sum_j w_{ij} \nu_j$ depends on the distribution of time-averaged firing rates

II. the time-averaged firing rate $\nu_i$ of a neuron should depend on its time-averaged synaptic drive $\bar{h}_i$

These two observations respectively suggest that we should be able to write down a set of *self-consistent* equations expressing

i. the moments of the synaptic drive $\langle \bar{h}^\ell \rangle$ in terms of the moments of the firing rate $\langle \nu^\ell \rangle$

ii. the moments of the firing rate $\langle \nu^\ell \rangle$ in terms of the moments of the synaptic drive $\langle \bar{h}^\ell \rangle$

Thus, for computing $L$ moments, we should be able to write down $2L$ equations with $2L$ unknowns - a system of equations we can (at least in principle) solve. This will require a few more assumptions along the way (namely, $N \to \infty$, i.i.d. weights, and a saturating gain function), but it will turn out to yield some interesting results.

We start with writing down equations expressing the moments of the distribution over time-averaged synaptic drives in terms of the moments of the distribution over time-averaged firing rates. We first note that the synaptic drive is a big sum of $N-1$ terms:

$$\bar{h}_i = \sum_j w_{ij} \nu_j$$

Thus, if each of the terms inside the sum are independently distributed over index $i$, as $N \to \infty$ (the regime we're interested in, since there's lots of neurons in the brain) the Central Limit Theorem tells us that the distribution of $\bar{h}_i$ over index $i$ (i.e. over the population of $N$ neurons) becomes Gaussian. Noting that the only terms on the right-hand-side that vary over index $i$ are the synaptic weights, we conclude that if the synaptic weights are independently sampled then the distribution over time-averaged synaptic drives in the population will approach a Gaussian in the large $N$ limit. For analytical convenience, we therefore incorporate the following structural constraint into our model:

> the synaptic weights $w_{ij}$ are independent and identically distributed (i.i.d.) with mean $\bar{w}$ and variance $\sigma_w^2$

While the assumption of i.i.d. weights may not hold in the brain, it might still allow our model to provide useful insights into how actual neural circuits operate[6].

We can now easily take the limit of large $N$ and invoke the CLT to write:

$$\bar{h}_i = \mu_h + \sqrt{N}\sigma_h \xi_i \quad \xi_i \sim \mathcal{N}(0,1)$$

where

$$\mu_h = \langle \bar{h}_i \rangle_i = \sum_j \langle w_{ij} \rangle_i \nu_j = \bar{w} \sum_j \nu_j = N\bar{w}\bar{\nu} \tag{3a}$$

$$\mathrm{Var}_i[\bar{h}_i] = \sum_j \mathrm{Var}_i[w_{ij}]\nu_j^2 = \sigma_w^2 \sum_j \nu_j^2 = N\sigma_w^2 \overline{\nu^2} \equiv N\sigma_h^2 \tag{3b}$$

where I have defined the $\ell$th moment of the firing rates as

$$\overline{\nu^\ell} \equiv \frac{1}{N} \sum_j \nu_j^\ell$$

and the $i$ subscript on the expectation and variance operators indicates an expectation over the random variation along index $i$. We also used the fact that, because the $w_{ij}$'s are identically distributed, $\forall j, k \ \langle w_{ij} \rangle_i = \langle w_{ik} \rangle_i = \bar{w}$ (and similarly for the second moment and therefore the variance).

Having written down a whole distribution over time-averaged synaptic drives $\bar{h}_i$ in terms of the first and second moments of the time-averaged firing rates $\nu_i$, we now turn to the problem of expressing the latter in terms of the former. This becomes easy once we have a way of transforming the time-averaged synaptic drive to a neuron into its time-averaged firing rate:

$$\nu_i = \phi_i\left(\bar{h}_i\right)$$

where the so-called "gain function" $\phi_i(\cdot)$ is some non-linear function depending on the single neuron dynamics $f_i(V_i, t)$ of neuron $i$. Generally, we'll take $\phi_i(\cdot) = \phi(\cdot)$ to simply be a scaled sigmoid saturating at $\nu_{max} \sim 100$Hz, reflecting the fact that firing rates cannot be negative and neurons can only fire at a finite rate. Further justification for the sigmoid shape is provided by Wilson & Cowan (1972): the time-averaged firing rate should be directly proportional to the probability of having suprathreshold input per unit time, approximately equal to the cumulative probability density of $\bar{h}_i$ from spiking threshold to infinity. If $\bar{h}_i$ has a unimodal distribution (which, as we just showed, it does under the assumption of i.i.d. weights), then this cumulative distribution will be a sigmoid[7]. Moreover, the sigmoid captures the main ingredients of what a neuronal gain function should look like: positive, monotonically increasing, and saturating. That said, most of the subsequent analysis is agnostic as to what the form of the gain function is, so any reasonable saturating function should do.

With this gain function in hand, we can easily express any moment of $\nu$ as a function of the mean $\mu_h$ and variance $N\sigma_h^2$ of the time-averaged synaptic drives $\bar{h}_i$. In the limit of $N \to \infty$, the law of large numbers give us:

I think that's the one

$$\overline{\nu^\ell} = \lim_{N\to\infty} \frac{1}{N} \sum_i \phi\left(\bar{h}_i\right)^\ell = \int \phi(\bar{h})^\ell P(\bar{h}) \, \mathrm{d}\bar{h} = \int \phi(\mu_h + \sqrt{N}\sigma_h \xi)^\ell \, \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}\xi \tag{4}$$

Putting equations 3 and 4 together gives us the following set of four self-consistent equations describing the mean and variance of $\bar{h}_i$ and $\nu_i$ over the population:

$$\begin{aligned}
\mu_h &= N\bar{w}\bar{\nu} \\
\sigma_h^2 &= \sigma_w^2 \overline{\nu^2} \\
\bar{\nu} &= \int \phi(\mu_h + \sqrt{N}\sigma_h \xi) \, \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}\xi \\
\overline{\nu^2} &= \int \phi(\mu_h + \sqrt{N}\sigma_h \xi)^2 \, \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}\xi
\end{aligned}$$

[6] Recall the famous UCL graduate statistician George Box: "all models are wrong; some models are useful"

[7] If $\bar{h}_i$ had a multimodal distribution, then $\phi(\cdot)$ would look like a stack of sigmoids, with an inflection point at each mode.

Numerically solving this system of equations gives us access to $\mu_h$ and $\sigma_h^2$, which allow us to compute any moment $\overline{\nu^\ell}$ of the time-averaged firing rates using equation 4.

We can now make a few simple observations about the distribution over time-averaged firing rates that will give some critical insights into how these depend on the distribution over synaptic weights $w_{ij}$. First, consider the case of $\text{Var}[\bar{h}_i] \to 0$. In this limit, $\overline{\nu^\ell} \to \phi^\ell(\mu_h)$, such that $\overline{\nu^2} = \bar{\nu}^2$ and $\text{Var}[\nu] = 0$, meaning that

$$P(\nu) \to \delta(\nu - \bar{\nu})$$

On the other hand, as $\text{Var}[\bar{h}_i] \to \infty$, $\nu_i = \phi(\bar{h}_i) \to \{0, \nu_{max}\}$, so that

$$P(\nu) \to \frac{1}{2}\delta(\nu) + \frac{1}{2}\delta(\nu - \nu_{max})$$

where $\nu_{max}$ is the maximum of the saturating gain function $\phi(\cdot)$ (it is easy to verify that this distribution has the exact moments given by equation 4, namely $\overline{\nu^\ell} = \frac{1}{2}\nu_{max}^\ell$). Given that actual observed firing rates are variable across neurons and generally lie somewhere between 0 and their maximum, these two results tell us that $\text{Var}[\bar{h}_i]$ should be greater than 0 and finite. We thus require that $\sigma_w^2 \sim 1/N$, so that in the limit of $N \to \infty$, $\text{Var}[\bar{h}_i] = N\sigma_w^2\overline{\nu^2} \sim \mathcal{O}(1)$. Furthermore, the mean time-averaged synaptic drive $\mu_h$ scales with $N$, meaning that in the large $N$ limit $\nu_i \to 0$ or $\nu_{max}$ (depending on whether $\bar{w}$ is negative or positive, respectively). We should thus also enforce that the mean weight $\bar{w} \sim 1/N$. This is our first result: in a randomly fully connected current-based network with sigmoidal gain functions, the mean and variance of the weights need to scale with $1/N$ for the dynamics not to saturate.

Another possibility could be to set $w_{ij} = \frac{\tilde{w}_{ij}}{N}$ with $\tilde{w}_{ij}$ i.i.d. with mean and variance $\tilde{w}, \sigma_{\tilde{w}^2} \sim \mathcal{O}(1)$, so that $\bar{w} = \frac{\tilde{w}}{N} \sim \mathcal{O}(1/N)$. However, this would entail $\sigma_w^2 = \frac{\sigma_{\tilde{w}}}{N^2} \Rightarrow \text{Var}[\bar{h}_i] \sim \mathcal{O}(\frac{1}{N})$, quickly leading to constant firing rates across the network (i.e. $P(\nu) = \delta(\nu - \bar{\nu})$) as $N \to \infty$. This problem illustrates the fine-tuning required to obtain realistic dynamics in a randomly fully connected network: it is hard to keep $\mu_h$ low while also ensuring $\text{Var}[\bar{h}_i] \sim \mathcal{O}(1)$. A viable alternative along these lines would be $w_{ij} = \frac{\tilde{w}_{ij}}{\sqrt{N}}$, in which case $\text{Var}[\bar{h}_i] \sim \mathcal{O}(1)$. But in this case $\mu_h \sim \mathcal{O}(\sqrt{N})$, thus only partially solving the problem since the mean synaptic drive will still (albeit slowly) tend towards the saturated regime as $N \to \infty$.

Such *dense* connectivity structure - in which every neuron is connected to every other neuron - evidently requires strong restrictions on the weights for the network to be able to generate realistic dynamics. This might also explain why it is rarely found in nature, where cortical connectivity rates, for example, are on the order of 10%. We can incorporate such *sparse* connectivity structure into our mean field equations by adding a parameter $K << N$ that controls the mean number of outgoing connections per neuron, so that the probability that any any two neurons are connected - called the *connectivity rate* - is equal to $K/N$ (because the network is randomly connected). We can then write

$$\begin{aligned}
w_{ij} &= \zeta_{ij}\tilde{w}_{ij} \\
\zeta_{ij} \in \{0,1\} &\sim \text{Bernoulli}\,(K/N) \\
\tilde{w}_{ij} &\text{ i.i.d. with mean } \tilde{w} \text{ and variance } \sigma_{\tilde{w}}^2 \\
\Rightarrow \mu_h &= N\bar{w}\bar{\nu} = N\frac{K}{N}\tilde{w}\bar{\nu} = K\tilde{w}\bar{\nu} \\
\Rightarrow \sigma_h^2 &= \sigma_w^2\overline{\nu^2} \\
&= \left(\langle\zeta^2\rangle\langle\tilde{w}^2\rangle - \langle\zeta\rangle^2\langle\tilde{w}\rangle^2\right)\overline{\nu^2} \\
&= \left(\frac{K}{N}(\sigma_{\tilde{w}}^2 + \tilde{w}^2) - \frac{K^2}{N^2}\tilde{w}^2\right)\overline{\nu^2} \\
&= \frac{K}{N}\left(\sigma_{\tilde{w}}^2 + \left(1 - \frac{K}{N}\right)\tilde{w}^2\right)\overline{\nu^2}
\end{aligned}$$

since $\zeta_{ij}, \tilde{w}_{ij}$ are independent. Writing it in this form makes it easy to interpret: in the large $N$ limit, the variance of the synaptic drive $\text{Var}[\bar{h}_i] = N\sigma_h^2$ is proportional to the variance of the (on average) $K$ non-zero input connections plus a correction for the remaining absent connections with weight 0. Thus, $\text{Var}[\bar{h}_i] \sim \mathcal{O}(K)$ is independent of $N$ regardless of $\sigma_{\tilde{w}}^2$. If we want to maintain the connectivity rate $p = K/N$ constant, however, $K \propto N$ so we will need $\tilde{w}, \sigma_{\tilde{w}} \sim 1/\sqrt{K}$ for

the firing rates to maintain variable non-saturating firing rates. However, it is still the case that $\mu_h \sim \mathcal{O}(\sqrt{K})$ so we will need either a really small $p$ or a really small constant of proportionality $k_h = \mu_h/\sqrt{(K)}$.

We can generalize the above analysis to the more realistic case of E/I networks that obey Dale's Law, where we have four types of synaptic weights $w_{ij}^{\mathrm{QR}} > 0$ from R neurons onto Q neurons, R,Q $\in$ {E,I} standing for excitatory (E) or inhibitory (I). The synaptic drive to a Q neuron is then given by

$$h_i^{\mathrm{Q}}(t) = \sum_{j\in E} w_{ij}^{\mathrm{QE}} g_j^{\mathrm{E}}(t) - \sum_{j\in I} w_{ij}^{\mathrm{QI}} g_j^{\mathrm{I}}(t) + I_{\mathrm{Q}}$$

where, for generality, I have also included an external constant input current $I_{\mathrm{Q}}$ injected equally into all $Q$ neurons. Assuming a fixed connectivity rate $p = K/N$, and setting

$$w_{ij}^{\mathrm{QR}} = \zeta_{ij}^{\mathrm{QR}} \frac{\tilde{w}_{ij}^{\mathrm{QR}}}{\sqrt{K}}$$
$$\zeta_{ij}^{\mathrm{QR}} \sim \text{Bernoulli}(p)$$
$$\tilde{w}_{ij}^{\mathrm{QR}} \text{ i.i.d. with mean } \bar{w}_{\mathrm{QR}} \sim \mathcal{O}(1) \text{ and variance } \sigma^2_{w_{\mathrm{QR}}} \sim \mathcal{O}(1)$$
$$I_{\mathrm{Q}} = \frac{\bar{I}_{\mathrm{Q}}}{\sqrt{K}}$$

we can exactly repeat the above derivation, giving us, in the large $N$ limit (for both subpopulations):

$$\overline{\nu_{\mathrm{Q}}^{\ell}} = \int \phi^{\ell}\left(\sqrt{K}\left(\bar{w}_{\mathrm{QE}}\overline{\nu}_{\mathrm{E}} - \bar{w}_{\mathrm{QI}}\overline{\nu}_{\mathrm{I}} + \bar{I}_{\mathrm{Q}}\right) + \sigma\xi\right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}\xi \tag{5a}$$

$$\sigma^2 = \sigma^2_{\mathrm{QE}} + \sigma^2_{\mathrm{QI}} \tag{5b}$$

$$\sigma^2_{\mathrm{QR}} = \left(\sigma^2_{w_{\mathrm{QR}}} + (1-p)\bar{w}^2_{\mathrm{QR}}\right)\overline{\nu^2_{\mathrm{R}}} \tag{5c}$$

In this case, we can hope for realistic dynamics even for unbounded $N$: if $\tilde{w}_{QE}, \tilde{w}_{QI}$ are picked carefully enough so that $\tilde{w}_{\mathrm{QE}}\overline{\nu}_{\mathrm{E}} - \tilde{w}_{\mathrm{QI}}\overline{\nu}_{\mathrm{I}} \approx 0$ (the so-called balanced regime), then, under infinitesimally small external input $\bar{I}_{\mathrm{Q}}$, $\sigma \sim \mathcal{O}(1)$ and we are guaranteed to stay within a brain-like regime of time-averaged firing rates.

In sum, for a current-based network model with

- saturating gain function $\phi(\cdot)$
- synaptic drive

$$h_i(t) = \sum_j w_{ij} g_j(t) + I_i^{ext}(t) = \underbrace{\sum_j w_{ij}\nu_j + I_i}_{\bar{h}_i} + \underbrace{\sum_j w_{ij}\delta g_j(t) + \delta I_i(t)}_{\delta h_i(t)}$$

- connectivity rate $p = K/N$
- random i.i.d. non-zero weights $w_{ij}$ with mean $\bar{w}$ and variance $\sigma_w^2$
- random i.i.d. time-averaged inputs $I_i \sim \mathcal{O}(\sum_j w_{ij}\nu_j)$ with mean $\bar{I}$ and variance $\sigma_I^2$

the following holds in the large $N \to \infty$ limit:

$$\boxed{\overline{\nu^{\ell}} = \int \phi\left(K\bar{w}\bar{\nu} + \bar{I} + \sqrt{K((\sigma_w^2 + (1-p)\bar{w}^2)\overline{\nu^2} + \sigma_I^2)}\xi\right)^{\ell} \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}\xi}$$

For such a network to have realistic dynamics (i.e. different firing rates across the population), it is therefore necessary that

$$\sigma_w^2 \sim \frac{1}{K}$$

We now briefly turn to problem #2 outlined above: characterizing temporal correlations in network activity. For this, we turn to the dynamic noise component of the synaptic drive, given by

$$\delta h_i(t) = \sum_j w_{ij} \delta g_j(t)$$

where $\delta g_j(t) \equiv g_j(t) - \bar{g}_j$ are 0-mean fluctuations of the input synaptic conductances. It turns out that with i.i.d. weights, in the limit of large $N$ two things happen: (i) the weights and fluctuations decouple, and (ii) the fluctuations across pairs of different neurons become uncorrelated. Intuitively, this is can be shown to be true when the connectivity is very sparse, when $K << N$ (i.e. $p << 1$, [Vreeswijk and Sompolinsky, 1998]). Somewhat less intuitively, it turns this is also true in E/I networks in the so-called "balanced state", where the total excitatory and inhibitory input to a given neuron are correlated and cancel each other out ([Renart et al., 2010, Rosenbaum et al., 2017])[8]. Thus, we again find ourselves with a big sum of independent random variables. In this case, however, we have a dynamical variable that is a function of time, so the CLT tells us that in the large $N$ limit $\delta h_i(t)$ becomes a draw from a Gaussian *process*.

Thus, all we need to fully characterize the statistics of the synaptic drive fluctuations $\delta h_i(t)$ is their mean and covariance. By construction, their mean is 0, and their covariance is captured by the cross-correlation function

$$C_{ij}(\tau) = \langle \delta h_i(t) \delta h_j(t+\tau) \rangle_t$$

where, following the above notational convention with expectations, the expectation is over time $t$. Given that in the large $N$ limit neurons are uncorrelated, $C_{ij} = 0$ whenever $i \neq j$. Furthermore, in our homogenous randomly connected network model, there is nothing to distinguish one neuron from another, so the only thing we really care about is the population mean autocorrelation function:

$$\begin{aligned}
C(\tau) &= \frac{1}{N} \sum_i C_{ii}(\tau) \\
&= \frac{1}{N} \sum_i \langle \delta h_i(t) \delta h_i(t+\tau) \rangle_t \\
&= \frac{1}{N} \sum_i \sum_{j,j'} w_{ij} w_{ij'} \langle \delta g_j(t) \delta g_{j'}(t+\tau) \rangle_t \\
&= \frac{1}{N} \sum_{i,j} w_{ij}^2 \langle \delta g_j(t) \delta g_j(t+\tau) \rangle_t \\
&= \sum_j \left( \frac{1}{N} \sum_i w_{ij}^2 \right) \langle \delta g_j(t) \delta g_j(t-\tau) \rangle_t \\
&\simeq N \overline{w^2} \Delta(\tau), \quad \Delta(\tau) = \frac{1}{N} \sum_j \langle \delta g_j(t) \delta g_j(t-\tau) \rangle_t
\end{aligned}$$

where the fourth equality follows from our assumption of no correlations, and the last approximation follows from the weights being i.i.d. and therefore self-averaging in the $N \to \infty$ limit.

We now have a full statistical characterization of the synaptic drives in a spiking network with i.i.d. weights (exact in the $N \to \infty$ limit under a few extra assumptions - namely, no correlations), as a function of the statistics of the spiking output (namely, the first and second moments of the time-averaged firing rates $\bar{\nu}, \overline{\nu^2}$ and the population mean autocorrelation of synaptic conductances $\Delta(\tau)$). However, unlike in the case of the quenched noise $\bar{h}_i$, it is not clear how to relate correlations in synaptic drives $C(\tau)$ to correlations in synaptic conductances $\Delta(\tau)$ to obtain a set of self-consistent equations we can solve for $C(\tau), \Delta(\tau)$. To my knowledge this has never been done analytically for a particular spiking neuron model $f_i(V_i, t)$ (except under the assumption of Poisson firing, which makes everything pretty straightforward given you know the f-I curve: [Grabska-Barwińska and Latham, 2014]).

However, in simplified rate-based models, the output autocorrelation $\Delta(\tau)$ can be solved analytically to provide substantial insight into the possible dynamical regimes of the network

[8] For a brief overview of the 20+ years it took for theorists to figure out why this was the case, see [Latham, 2017]

([Sompolinsky et al., 1988, Mastrogiuseppe and Ostojic, 2017]). For example, consider a randomly connected *tanh* network with dynamics

$$\dot{x}_i = -x_i + \sum_{j=1}^{N} w_{ij}\phi(\gamma x_j)$$

with i.i.d. 0-mean weights $w_{ij} \sim \mathcal{N}(0, 1/N)$. In this case, the "synaptic conductances" are given by $g_j(t) = \phi(\gamma x_j(t))$, where $\gamma$ parametrizes the steepness of the non-linearity $\phi(\cdot) = \tanh(\cdot)$. Given this simple relationship between the synaptic conductances and the neurons' "potentials" $x_i$, one can show through some tedious but straightforward algebra that the autocorrelation satisfies the following differential equation:

$$\Delta(\tau) - \frac{\mathrm{d}^2\Delta}{\mathrm{d}\tau^2} = C(\tau)$$

with the following intuitive boundary conditions:

- $\Delta(\tau) \leq \Delta(0)$ since the autocorrelation will always be highest at 0 time lag
- $\Rightarrow \left.\frac{\mathrm{d}\Delta}{\mathrm{d}\tau}\right|_{\tau=0} = 0$ since $\tau = 0$ is an extremum of the function
- $\Rightarrow \left.\frac{\mathrm{d}^2\Delta}{\mathrm{d}\tau^2}\right|_{\tau=0} < 0$ since $\tau = 0$ is a maximum

Along with the above ODE, these boundary conditions allow you to make general statements about the shape of $\Delta(\tau)$, which gives qualitative insights into the trajectories of the system. Particularly, for certain values of $\gamma$, $\Delta(\tau) \to 0$ as $\tau \to \infty$, indicating that trajectories diverge over time and the system is chaotic ([Sompolinsky et al., 1988]).

## 3.2 Wilson-Cowan Equations

Given our above equations for the means of the distribution over time-averaged firing rates $\nu_{\mathrm{Q}}$, we might then use them to model the macroscopic dynamics of the population. Since $\overline{\nu}_{\mathrm{E}}, \overline{\nu}_{\mathrm{I}}$ depend on each other, given some initial condition or perturbation they will evolve over time until settling at an equilibrium, i.e. a solution to the pair of equations given by equation 5 for Q = E,I. We can thus model these dynamics by writing a simple differential equation for each subpopulation mean time-averaged firing rate $\overline{\nu_{\mathrm{Q}}}$, with equilibrium given by the right-hand-side of equation 5:

$$\tau_{\mathrm{E}}\dot{\nu}_{\mathrm{E}} = \psi_{\mathrm{E}}(\nu_{\mathrm{E}}, \nu_{\mathrm{I}}) - \nu_{\mathrm{E}}$$
$$\tau_{\mathrm{I}}\dot{\nu}_{\mathrm{I}} = \psi_{\mathrm{I}}(\nu_{\mathrm{E}}, \nu_{\mathrm{I}}) - \nu_{\mathrm{I}}$$

where, as per equation 5,

$$\psi_{\mathrm{Q}}(\nu_{\mathrm{E}}, \nu_{\mathrm{I}}) = \int \phi_{\mathrm{Q}}\left(\sqrt{K}\left(\bar{w}_{\mathrm{QE}}\overline{\nu}_{\mathrm{E}} - \bar{w}_{\mathrm{QI}}\overline{\nu}_{\mathrm{I}}\right) + I_{\mathrm{Q}} + \sigma\xi_i^Q\right) \frac{e^{-\frac{\xi^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}\xi$$

(assuming the same mean number of outgoing connections for each subpopulation, athough one could also easily carry out the derivation of 5 but for different subpopulation connectivity rates $K_{\mathrm{Q}}$). Thus, $\psi_{\mathrm{Q}}(\nu_{\mathrm{E}}, \nu_{\mathrm{I}})$ is just a Gaussian-smoothed version of the corresponding individual neuron gain function $\phi_{\mathrm{Q}}(\cdot)$. In the case of sigmoidal sigmoidal $\phi_{\mathrm{Q}}(\cdot)$, then, $\psi_{\mathrm{Q}}(\cdot)$ will be sigmoidal as well. Note that we have explicitly included constant external inputs $I_{\mathrm{E}}, I_{\mathrm{I}}$ to each subpopulation.

These are called the Wilson-Cowan equations, and can be alternatively derived from very general assumptions (Wilson & Cowan, 1972). We now want to analyze this two-dimensional dynamical system to try to understand the following experimental observations of the mammalian neocortex:

1. low time-averaged firing rates ($\sim$ .2Hz)
2. network oscillations
3. UP and DOWN states of high and low average membrane potential, respectively, which are particularly common under anesthesia and last on the order of seconds

4. bumps of membrane potential that last on the order of 10's of milliseconds, separated by DOWN states on the order of seconds

It turns out our simplified model of time-averaged population mean activity can produce all four of these patterns, under certain regimes.

To see this, we turn to a stability analysis of our two-dimensional system to get an idea of its qualitative behavior. The nullclines of this system are given by

$$\nu_{\mathrm{E}} = \psi_{\mathrm{E}}(\nu_{\mathrm{E}}, \nu_{\mathrm{I}})$$
$$\nu_{\mathrm{I}} = \psi_{\mathrm{I}}(\nu_{\mathrm{E}}, \nu_{\mathrm{I}})$$

We can draw these nullclines by plotting $\psi_{\mathrm{Q}}$ as a function of $\nu_{\mathrm{Q}}$ and finding where it cross the unity line $\nu_{\mathrm{Q}} = \psi_{\mathrm{Q}}(\nu_{\mathrm{Q}}, \nu_{\mathrm{R}})$, for different values of $\nu_{\mathrm{R}}$. By plotting all of these intersections in the $\nu_{\mathrm{I}} - \nu_{\mathrm{E}}$ plane at their corresponding values of $(\nu_{\mathrm{E}}, \nu_{\mathrm{I}})$, we get the $\nu_{\mathrm{Q}}$-nullcline. This is done in figure 4, with $Q = E$ in the top graph (A) and $Q = I$ in the bottome left graph (B), giving us the $\nu_{\mathrm{E}}$ and $\nu_{\mathrm{I}}$ nullclines in black and grey, respectively, in the bottom right graph (C). Here a generic pair of sigmoidal $\psi_{\mathrm{E}}, \psi_{\mathrm{I}}$ functions are used, with $I_{\mathrm{E}} = I_{\mathrm{I}} = 0$.

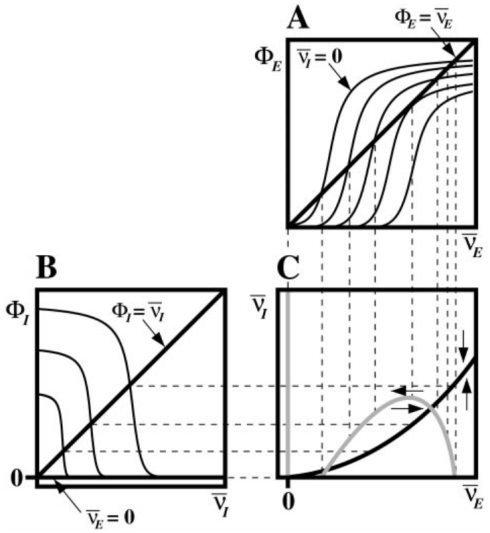



Figure 4: Copied from Latham et al. (2000). $\phi_{\mathrm{Q}} = \psi_{\mathrm{Q}}$ on the labels

We now proceed to analyze each of the fixed points, given by the three intersections of the nullclines in figure 4C. It turns out we can do a lot geometrically, without explicitly parametrizing any of the equations above. Let $(\nu_{\mathrm{E}}^*, \nu_{\mathrm{I}}^*)$ designate the location of one such fixed point, such that

$$\nu_{\mathrm{E}}^* = \psi_{\mathrm{E}}(\nu_{\mathrm{E}}^*, \nu_{\mathrm{E}}^*)$$
$$\nu_{\mathrm{I}}^* = \psi_{\mathrm{I}}(\nu_{\mathrm{I}}^*, \nu_{\mathrm{E}}^*)$$

As per standard stability analysis (section 7.5), we can find out if this fixed point is stable by looking at the Jacobian matrix of the dynamical system evaluated at the fixed point, here given by

$$\mathbf{J} = \begin{bmatrix} \tau_{\mathrm{E}}^{-1}(\psi_{\mathrm{E,E}} - 1) & \tau_{\mathrm{E}}^{-1}\psi_{\mathrm{E,I}} \\ \tau_{\mathrm{I}}^{-1}\psi_{\mathrm{I,E}} & \tau_{\mathrm{I}}^{-1}(\psi_{\mathrm{I,I}} - 1) \end{bmatrix}$$

where we define

$$\psi_{\mathrm{Q,R}} \equiv \left.\frac{\partial \psi_{\mathrm{Q}}}{\partial \nu_{\mathrm{R}}}\right|_{\nu_{\mathrm{E}}^*, \nu_{\mathrm{I}}^*}$$

We then know that the fixed point will be stable iff the trace and determinant of $\mathbf{J}$ are negative and positive, respectively, giving us the following conditions for the stability of $(\nu_{\mathrm{E}}^*, \nu_{\mathrm{I}}^*)$:

$$\frac{1 - \psi_{\mathrm{I,I}}}{\tau_{\mathrm{I}}} > \frac{\psi_{\mathrm{E,E}} - 1}{\tau_{\mathrm{E}}}$$
$$(\psi_{\mathrm{E,E}} - 1)(\psi_{\mathrm{I,I}} - 1) > \psi_{\mathrm{E,I}}\psi_{\mathrm{I,E}}$$

It turns out we can verify these conditions geometrically by expressing the slopes of the nullclines in terms of the partial derivatives $\psi_{\text{Q,R}}$. To do so, we consider a perturbation from the fixed point $(\nu_{\text{E}}^*, \nu_{\text{I}}^*) \to (\nu_{\text{E}}^* + \delta_{\text{E}}\nu_{\text{E}},\, \nu_{\text{I}}^* + \delta_{\text{E}}\nu_{\text{I}})$ *along the* $\nu_E$*-nullcline* - hence the $E$ subscript on the $\delta$'s. Since we know the equation for the nullcline, we can use this to compute its slope at the fixed point via a first-order Taylor approximation:

$$\begin{aligned}
\nu_{\text{E}}^* + \delta_{\text{E}}\nu_{\text{E}} &= \psi_{\text{E}}(\nu_{\text{E}}^* + \delta_{\text{E}}\nu_{\text{E}},\, \nu_{\text{I}}^* + \delta_{\text{E}}\nu_{\text{I}}) \\
&\approx \psi_{\text{E}}(\nu_{\text{E}}^*, \nu_{\text{I}}^*) + \delta_{\text{E}}\nu_{\text{E}} \frac{\partial \psi_{\text{E}}}{\partial \nu_{\text{E}}}\bigg|_{\nu_{\text{E}}^*, \nu_{\text{I}}^*} + \delta_{\text{E}}\nu_{\text{I}} \frac{\partial \psi_{\text{E}}}{\partial \nu_{\text{I}}}\bigg|_{\nu_{\text{E}}^*, \nu_{\text{I}}^*} \\
&= \nu_{\text{E}}^* + \delta_{\text{E}}\nu_{\text{E}}\psi_{\text{E,E}} + \delta_{\text{E}}\nu_{\text{I}}\psi_{\text{E,I}} \\
\Leftrightarrow \delta_{\text{E}}\nu_{\text{E}} &= \delta_{\text{E}}\nu_{\text{E}}\psi_{\text{E,E}} + \delta_{\text{E}}\nu_{\text{I}}\psi_{\text{E,I}} \\
\Leftrightarrow \frac{\delta_{\text{E}}\nu_{\text{I}}}{\delta_{\text{E}}\nu_{\text{E}}} &= \frac{1 - \psi_{\text{E,E}}}{\psi_{\text{E,I}}}
\end{aligned}$$

Performing the same calculation for the $\nu_{\text{I}}$-nullcline, we have that the slopes of the excitatory and inhibitory nullclines at the fixed point are given by, respectively

$$\begin{aligned}
m_{\text{E}} &= \frac{\delta_{\text{E}}\nu_{\text{I}}}{\delta_{\text{E}}\nu_{\text{E}}} = \frac{1 - \psi_{\text{E,E}}}{\psi_{\text{E,I}}} \\
m_{\text{I}} &= \frac{\delta_{\text{I}}\nu_{\text{I}}}{\delta_{\text{I}}\nu_{\text{E}}} = \frac{\psi_{\text{I,E}}}{1 - \psi_{\text{I,I}}}
\end{aligned}$$

Because $\psi_{\text{Q,E}} > 0, \psi_{\text{Q,I}} < 0$ we then know that

$$\begin{aligned}
&m_{\text{I}} > 0 \text{ always} \\
&m_{\text{E}} \begin{cases} < 0 \Leftrightarrow 0 < \psi_{\text{E,E}} < 1 \\ > 0 \Leftrightarrow \psi_{\text{E,E}} > 1 \end{cases}
\end{aligned}$$

We can now relate our stability conditions on the partial derivatives $\psi_{\text{Q,R}}$ to statements about the nullcline slopes at the fixed point:

$$\begin{aligned}
m_{\text{E}} < m_{\text{I}} &\Leftrightarrow (\psi_{\text{E,E}} - 1)(\psi_{\text{I,I}} - 1) > \psi_{\text{E,I}}\psi_{\text{I,E}} \\
m_{\text{E}} < 0 &\Rightarrow \frac{1 - \psi_{\text{I,I}}}{\tau_{\text{I}}} > \frac{\psi_{\text{E,E}} - 1}{\tau_{\text{E}}}
\end{aligned}$$

The first statement tells us that for the fixed point $(\nu_{\text{E}}^*, \nu_{\text{I}}^*)$ to be stable, the slope of the inhibitory nullcline at this point has to be steeper than that of the excitatory nullcline. This is intuitive: the inhibitory population mean firing rate $\nu_{\text{I}}$ has to be more sensitive to changes in the excitatory population mean firing rate than the excitatory population itself for the negative feedback to be strong enough to generate a stable state. Graphically, this translates to the inhibitory nullcline intersecting the excitatory nullcline at $(\nu_{\text{E}}^*, \nu_{\text{I}}^*)$ from below.

The second condition tells us that if the excitatory nullcline has negative slope at $(\nu_{\text{E}}^*, \nu_{\text{I}}^*)$ (i.e. the fixed point lies on a *stable branch* of the $\nu_{\text{E}}$-nullcline), we know that the fixed point is stable, otherwise (the fixed point lies on an *unstable branch* of the $\nu_{\text{E}}$-nullcline) we don't know - it depends on the values of $\psi_{\text{E,E}}, \psi_{\text{I,I}}, \tau_{\text{E}}, \tau_{\text{I}}$. Particularly, given similar $\tau_{\text{E}} \approx \tau_{\text{I}}$, we can see from our original stability conditions that the fixed point will be stable for highly negative $\psi_{\text{I,I}}$ and small $\psi_{\text{E,E}}$. In other words, if $m_{\text{E}} > 0$ at the fixed point, then the fixed point will only be stabile if there is weak coupling between excitatory firing rates - i.e. weaker positive feedback - and strong coupling between inhibitory neurons - i.e. stronger disinhibiton. This latter requirement might seem counterintuitive, but it makes sense mathematically when you consider that it is the only coupling able to pull local perturbations from an equilibrium back to it: $E$-$E$ coupling repels them further away and $E$-$I$,$I$-$E$ couplings rotates them around the equilibrium, but $I$-$I$ coupling brings them back. Thus, a strong $I$-$I$ coupling - i.e. a highly negative $\psi_{\text{I,I}}$ - is necessary for a fixed point on an unstable branch of the $\nu_{\text{E}}$-nullcline to be stabile. Note that in the case of $m_{\text{E}} > 0$, a small $\psi_{\text{E,E}}$ implies a small $m_{\text{E}}$, i.e. a more shallow positive slope (recall that $\psi_{\text{E,I}} < 0$ so in this case $\psi_{\text{E,E}} \geq 1$). So as the excitatory nullcline slope gets larger at the fixed point, it becomes more unstable (eventually leading to a Hopf bifurcation).

Using these conditions, we can now go back to our standard $E$-$I$ nullclines in figure 4C and conclude that the left and right intersections correspond to stable fixed points (with $m_{\text{E}} < m_{\text{I}}, m_{\text{E}} < 0$) and the fixed point in between is therefore unstable (thus separating their basins of attraction). Thus, this set of nullclines is inconsistent with the observation of low average firing rates in cortex. While the rightmost stable state is at a high firing rate, the leftmost one is at (0,0), corresponding to the state where no neurons are firing and thus no neurons can or will fire (thus making it stable). To get a stable state at a low firing rate, one can see graphically that we need the excitatory and inhibitory nullclines to both shift upwards such that the leftmost intersection is pushed out of the origin. Recalling that $\psi_{\text{E}}(\nu_{\text{E}}, \nu_{\text{I}}), \psi_{\text{I}}(\nu_{\text{E}}, \nu_{\text{I}})$ are monotonically increasing functions of $I_{\text{E}}, I_{\text{I}}$, we can shift the nullclines up by simply increasing the external current inputs $I_{\text{E}}, I_{\text{I}}$, which until now were set to 0. This corresponds to shifting the $\psi_{\text{E}}, \psi_{\text{I}}$ curves in figure 4A,B to the left and right, respectively. It turns out that if we increase $I_{\text{E}}, I_{\text{I}}$ enough so that $\psi_{\text{E}}(0,0), \psi_{\text{I}}(0,0) > 0$ (i.e. shift the $\psi_{\text{E}}, \psi_{\text{I}}$ curves until the $y$-intercept is above 0, see regime 3 in figure 2 of Latham et al., 2000), we end up with nullclines roughly looking like those in figure 5, with an equilibrium at low mean firing rates. Biologically, this shift in the equilibrium gain functions such that their $y$-intercepts are above 0 can be interpreted as there being a number of *endogenously active* cells in the population, which can have non-zero firing rates in the abscence of any recurrent input from the rest of the population. Our analysis thus suggests that an external current input strong enough to endow each subpopulation with endogenously active cells is necessary for networks to have a stable equilibrium at low average firing rates, like what is observed in the mammalian neocortex (Latham et al. 2000). In a real brain, this external current could come from upstream inputs, membrane potential noise, or some single-cell intrinsic membrane currents.

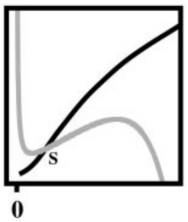



Figure 5: Copied from Latham et al. (2000). Axes as in figure 4C.

But we still have to confirm that the new equilibrium we have found is stable. As drawn in the figure, the equilibrium lies on the unstable branch of the excitatory nullcline (i.e. $m_{\text{E}} > 0$), meaning that its stability depends on the strength of the $E$-$E$ and $I$-$I$ coupling. We could instead push the inhibitory nullcline further up so that it intersects the excitatory one on its leftmost stable branch, but this would create a stable equilibrium where the mean inhibitory firing rate is greater than the mean excitatory firing rate. Such a regime is rarely seen in real brains, and it can be shown that the minimum of the excitatory nullcline is so small ($\sim$ .01Hz) that this equilibrium would correspond to unrealistically low excitatory firing rates (Latham et al., 2000). We thus conclude that the equilibrium has to be on the unstable branch of the excitatory nullcline as in figure 5, so its stability depends on the $E$-$E$ and $I$-$I$ coupling.

We can regulate this coupling by changing the mean synaptic weight strengths $\bar{w}_{\text{EE}}, \bar{w}_{\text{II}}$. But note that this also results in shifts of the nullclines: namely, increasing excitatory synaptic weights $\bar{w}_{\text{EE}}(\bar{w}_{\text{IE}})$ leads to upward shifts of the excitatory(inhibitory) nullclines while increasing inhibitory synaptic weights $\bar{w}_{\text{EI}}(\bar{w}_{\text{II}})$ lowers them. Since a rise in the excitatory/inhibitory nullcline pushes the equilibrium to higher/lower mean firing rates, this translates to high $\bar{w}_{\text{EE}}, \bar{w}_{\text{II}}$ pushing the equilibrium mean firing rates up and high $W_{\text{EI}}, W_{\text{IE}}$ pushing them down. Summing this all up, a randomly connected network with endogenously active cells will have a stable equilibrium with low mean firing rates (observation #1) if the mean $E$-$E$ connection strenghts are weak (to stabilize and lower the equilibrium), the mean $E$-$I$ and $I$-$E$ connection strenghts are strong (to lower the equilibrium), and the mean $I$-$I$ connection strengths are strong (to stabilize) but not too strong (to lower).

On the other hand, if the connectivity is such that the $E$-$E$/$I$-$I$ connections are too strong/weak, then the mean firing rate equilibrium becomes unstable. But we now note that the derivatives on the boundaries of figure 5 all point inwards: anything above(below) the $\nu_{\text{I}}$-nullcline has a downward(upward) facing derivative (since $\nu_{\text{I}} > (<) \psi_{\text{I}}(\nu_{\text{E}}, \nu_{\text{I}})$), and anything left(right) of the $\nu_{\text{E}}$-

nullcline has a rightward(leftward) facing derivative (since $\nu_{\text{E}} < (>)\, \psi_{\text{I}}(\nu_{\text{E}}, \nu_{\text{I}})$). By the Poincarè-Bendixson Theorem (footnote 4), then, there must be a limit cycle around this unstable equilibrium. In other words, a randomly connected network with strong $E$-$E$ connections and weak $I$-$I$ conenctions can exhibit oscillations, like those observed in the cortex (observation #2). This transition from stable to unstable + oscillations is called a *Hopf bifurcation.*

Finally, we consider the more biologically realistic case of dynamics under spike-frequency adaptation. We consider a parameter regime where there are few endogenously active cells, such that the nullclines look like those in figure 6. We incorporate spike-frequency adaptation into our model simply by expressing the external current input $I_{\text{Q}}$ as a dynamical variable:

$$I_{\text{Q}} = \theta_{\text{Q}} - g_{\text{Q}}$$
$$\tau_{\text{SFA}} \frac{\mathrm{d}g_{\text{Q}}}{\mathrm{d}t} = G_{\text{Q}}\nu_{\text{Q}} - g_{\text{Q}}$$

where $G_{\text{Q}}$ is a constant, $\theta_{\text{Q}}$ reflects the number of endogenously active cells under no adaptation (in the form of some kind of external current), and $Q \in \{E, I\}$. Thus, as the mean firing rate of subpopulation $Q$ increases, $g_{\text{Q}}$ increases, pushing $I_{\text{Q}}$ below 0 as the endogenously and non-endogenously active cells are silenced via spike-frequency adaptation. This results in a downward shift of the nullclines, as illustrated in figures 6 A→B→C. In this parameter regime where $\theta_{\text{Q}}$ is not too big, this results in a series of bifurcations: as the nullclines shift from A to B, a new 0-firing rate equilibrium is created, to which the network is forced to in the shift from B to C as the original non-zero firing rate equilibrium is destroyed. Once the firing rates drop to this new equilibrium, $g_{\text{Q}}$ goes to 0 and $I_{\text{Q}} \to \theta_{\text{Q}}$, pushing the equilibrium back up to its non-adapted state. Crucially, $\tau_{\text{SFA}}$, on the order of seconds, is much larger than $\tau_{\text{E}}, \tau_{\text{I}}$, so the network settles to equilibrium between the bifurcations. This leads to bursting: transitions between states of high mean firing rates (UP states) and states of very low mean firing rates (DOWN states), which last on the order of seconds (namely, on the order of spike-frequency adaptiation time $\tau_{\text{SFA}}$). We thus see that, in this regime, randomly connected networks can replicate observation #3. (Latham et al., 2000)

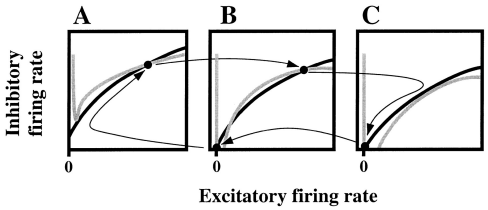



Figure 6: Copied from Latham et al. (2000).

Finally, we note that the nullclines in figure 6B can also generate observation #4, in the abscence of spike-frequency adaptation. In this case, we have a stable equilibrium at 0 mean firing rate and a stable or unstable equilibrium at higher firing rate separated by a saddle. If this higher equilibrium is unstable, then any perturbation to the lower stable equilibrium that pushes it rightward beyond the saddle will lead to a trajectory that loops around the high firing rate equilibrium before returning to the 0 mean firing rate equilibrium. This is evocative of bump responses, where perturbation or stimulus-evoked responses lead to a short period of high membrane potential (and thus high firing rates) before quickly returning to a DOWN state (observation #4).

## 3.3 Hopfield Network

We consider a fully-connected network of $N+1$ neurons with discrete states given by $s_i(t) \in \{1, -1\}$. We model their dynamics in discrete time, using the update

$$s_i(t+1) = \text{sign}\left[\sum_{j=1}^{N} W_{ij} s_j(t)\right]$$

where

$$\text{sign}[x] = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{else} \end{cases}$$

The crucial property of the Hopfield network is its connectivity matrix, given by

$$W_{ij} = \frac{1}{N}\sum_{m=1}^{M} \xi_i^{(m)}\xi_j^{(m)}, \quad W_{ii} = 0$$
$$\xi_i^{(m)} = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

All connections are thus symmetric $W_{ij} = W_{ji}$, with no autapses.

Suppose now that $\forall i\ s_i(t) = \xi_i^{(m')}$. We then have

$$\begin{aligned}
s_i(t+1) &= \text{sign}\left[\sum_{j\neq i} W_{ij}s_j(t)\right] \\
&= \text{sign}\left[\sum_{j\neq i}\frac{1}{N}\sum_{m=1}^{M}\xi_i^{(m)}\xi_j^{(m)}\xi_j^{(m')}\right] \\
&= \text{sign}\left[\frac{1}{N}\sum_{j\neq i}\left(\xi_i^{(m')}\xi_j^{(m')}\xi_j^{(m')} + \sum_{m\neq m'}\xi_i^{(m)}\xi_j^{(m)}\xi_j^{(m')}\right)\right] \\
&= \text{sign}\left[\frac{1}{N}\sum_{j\neq i}\left(\xi_i^{(m')} + \sum_{m\neq m'}\xi_i^{(m)}\xi_j^{(m)}\xi_j^{(m')}\right)\right] \\
&= \text{sign}\left[\xi_i^{(m')} + \frac{1}{N}\sum_{j\neq i}\sum_{m\neq m'}\xi_i^{(m)}\xi_j^{(m)}\xi_j^{(m')}\right] \\
&= \text{sign}\left[\xi_i^{(m')} + \eta_i\right]
\end{aligned}$$

Since each of the terms inside the sum are independently distributed, in the limit of large $N$ we can use the CLT to approximate $\eta_i$ with a Gaussian random variable:

$$\eta_i \to \mu + \frac{\sigma}{\sqrt{N}}\zeta_i, \quad \zeta_i \sim \mathcal{N}(0,1)$$
$$\mu = \left\langle \sum_{m\neq m'}\xi_i^{(m)}\xi_j^{(m)}\xi_j^{(m')}\right\rangle = \sum_{m\neq m'}\langle\xi_i^{(m)}\rangle\langle\xi_j^{(m)}\rangle\langle\xi_j^{(m')}\rangle = 0$$
$$\sigma^2 = \sum_{m\neq m'}\left\langle {\xi_i^{(m)}}^2\right\rangle\left\langle {\xi_j^{(m)}}^2\right\rangle\left\langle {\xi_j^{(m')}}^2\right\rangle = \sum_{m\neq m'}(1)(1)(1) = M-1$$

Thus, we have:

$$s_i(t+1) = \text{sign}\left[\xi_i^{(m')} + \sqrt{\frac{M-1}{N}}\zeta_i\right], \quad \zeta_i \sim \mathcal{N}(0,1)$$

In other words, as long as $M << N$, $s_i(t) = \xi_i^{(m')}$ is an equilibrium of the system, since $s_i(t) \approx s_i(t+1)$.

Generalizing the above for all possible $m'$, the system has $M$ such equilibria, each being a local minimum of the network's *energy* given by

$$E = -\frac{1}{2}\sum_{i,j} W_{ij}s_i s_j$$

The Hopfield network thus implements a kind of associative memory, whereby feeding it some input leads it to converge to the equilibrium state - or "memory" - most similar to the input. The Hopfield network can store $M$ such memories as long as it has many more neurons than memories $N >> M$, namely at least 1000 neurons for every 138 memories (i.e. $\frac{M}{N} \leq 0.138$; Amit, Gutfreund & Sompolinsky, 1987).

# 4 Functional Models of Synaptic Plasticity

Aside from NMDA-mediated plasticity, long-term plasticity is generally not well understood and thus generally modelled more phenomenologically. Rather than thinking about neurotransmitter release probabilities and neurotransmitter receptor conductances, here we directly model changes to the "weight" $W_{ij} = \bar{g}_{ij} P_{rel}^{(ij)}$ of a synapse from pre-synaptic neuron $j$ onto post-synaptic neuron $i$. More specifically, we will model these changes as a function of pre- and post- synaptic activity:

$$\tau_w \frac{\mathrm{d}W_{ij}}{\mathrm{d}t} = f_{ij}(r_j, r_i)$$

with $r_j, r_i$ for the pre- and post- synaptic cell activity (e.g. firing rate), respectively. $\tau_w$ sets the timescale of synaptic plasticity, analagous to a learning rate (large $\tau_w \rightarrow$ small learning rate).

Experimentally, a proxy for measuring $W_{ij}$ is the change in membrane potential in a post-synaptic neuron induced by triggering an action potential in a pre-synaptic neuron, called the *post-synaptic potential (PSP) amplitude*. A classic experimental setup consists of inducing high-frequency ($\sim$ 100Hz) bursts of action potentials simultaneously in both the pre- and post- synaptic neurons and then measuring the PSP amplitude. Such a stimulation protocol will usually lead to a jump in the PSP amplitude that can last on the order of hours. However, if you block protein synthesis, the PSP amplitude decays back to its baseline level before the burst protocol soon afterwards. We thus distinguish between "early" and "late" *long-term potentiation (LTP)*, where late LTP requires protein synthesis and early LTP does not[9]. Similarly, *long-term depression (LTD)* can be triggered via a low-frequency ($\sim$ 2Hz) bursting protocol. D&A fig. 8.1 shows the classic picture of this, where LFP amplitude in a hippocampal slice was used as a proxy for PSP amplitude (note the initial transient increase after the high frequency protocol, reflecting early LTP).

We now turn to phenomenological models of synaptic plasticity that can produce both LTD and LTP, as well as an array of other experimental observations. To facilitate analysis, we begin by considering a single post-synaptic neuron with linear dynamics:

$$\tau \frac{\mathrm{d}v}{\mathrm{d}t} = -v + \mathbf{w}^{\mathrm{T}}\mathbf{u}$$

For the rest of this section I adopt the convention of $v$ and $u_i$ for pre- and post- synaptic neuron activity. Importantly, we additionally assume that synaptic plasticity occurs on a much slower timescale than the neural dynamics (i.e. $\tau << \tau_w$), such that we can assume that the post-synaptic activity has converged to its equilibrium

$$v = \mathbf{w}^{\mathrm{T}}\mathbf{u}$$

Later we consider feed-forward and recurrent networks of neurons.

## 4.1 Hebb Rule

One interpretation of the above observation is the occurence of Hebbian plasticity: "neurons that fire together wire together." Since the high-frequency stimulation is likely to induce post-synaptic firing, synapses should strengthen because of co-occurence of high pre- and post- synaptic activity. Conversely, low-frequency stimulation is less likely to induce post-synaptic spikes, leading to weakening of the synapses and thus LTD. This gives us the basic *Hebb learning rule*:

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = v\mathbf{u}$$

Averaging over pre-synaptic inputs from trial to trial (e.g. over a lifetime) and taking the equilibrium post-synaptic activity $v = \mathbf{w}^{\mathrm{T}}\mathbf{u} = \mathbf{u}^{\mathrm{T}}\mathbf{w}$, we have:

$$\tau_w \left\langle \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} \right\rangle = \langle \mathbf{u}v \rangle = \langle \mathbf{u}\mathbf{u}^{\mathrm{T}} \rangle \mathbf{w} = \mathbf{Q}\mathbf{w}$$

where $\mathbf{Q}$ is the correlation matrix over pre-synaptic activity input patterns. We thus call the Hebb learning rule a *correlation-based plasticity rule*.

[9] This has lead to the *synaptic tagging hypothesis*, which postulates that simultaneous pre- and post- synaptic activity leads to the "tagging" of synapses, signalling the cell to insert more receptors there.

It is easy to see that the Hebb rule will lead to LTP whenever pre- and post- synaptic activity is correlated, but it can't lead to LTD if the activity vectors are positive. For this, we introduce an activity threshold:

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = (v - \theta_v)\mathbf{u}$$

or

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = v(\mathbf{u} - \theta_\mathbf{u})$$

If the pre- or post- synaptic activity is now below its corresponding threshold, the derivative of $\mathbf{w}$ becomes negative and LTD occurs. A natural setting for this threshold is the mean background firing activity $\theta_v = \langle v \rangle, \theta_\mathbf{u} = \langle \mathbf{u} \rangle$. Again averaging over all inputs over time, the result of adding a threshold gives us a *covariance-based plasticity rule*:

$$\tau_w \left\langle \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} \right\rangle = \langle \mathbf{u}(v - \langle v \rangle) \rangle = \langle \mathbf{u} \left( \mathbf{u}^\mathrm{T}\mathbf{w} - \langle \mathbf{u}^\mathrm{T} \rangle \mathbf{w} \right) \rangle = \left( \langle \mathbf{u}\mathbf{u}^\mathrm{T} \rangle - \langle \mathbf{u} \rangle \langle \mathbf{u} \rangle^\mathrm{T} \right) \mathbf{w} = \mathbf{C}\mathbf{w}$$

where $\mathbf{C}$ is the covariance matrix of the input patterns. One can easily verify that using a pre-synaptic activity threshold $\theta_\mathbf{u} = \langle \mathbf{u} \rangle$ instead of a post-synaptic activity threshold leads to the same average dynamics. However, having one or the other threshold leads to entirely different biological predictions:

- If only a post-synaptic activity threshold $\theta_v$ is included, then plasticity is induced at the $i$th synapse (i.e. $\mathrm{d}w_i/\mathrm{d}t \neq 0$) iff there is pre-synaptic activity at that synapse (i.e. $u_i > 0$). This is called *homosynaptic plasticity.*
- If only a pre-synaptic activity threshold $\theta_\mathbf{u}$ is included, then plasticity is induced at the $i$th synapse whenever there is post-synaptic activity (i.e. $v > 0$), even if there is no pre-synaptic activity at that synapse (i.e. $u_i = 0$). This leads to *heterosynaptic plasticity.*
- If both thresholds are included, then the model counterintuitively predicts that there should be LTP when both pre- and post- synaptic activity are below threshold

Because the weight dynamics are linear, we can easily analyze the result of applying these simple Hebbian synaptic learning rules. We first note that because $\mathbf{C}$ is symmetric, its eigenvectors $\mathbf{e}_i$ are orthogonal and form a complete basis for the space of $\mathbf{w}$. We can thus write

$$\mathbf{w}(t) = \sum_i c_i(t)\mathbf{e}_i$$

where the coefficients are simply equal to the scalar projection of $\mathbf{w}$ onto each eigenvector, given simply by $c_i(t) = \mathbf{w}^\mathrm{T}\mathbf{e}_i$ when we assume the eigenvectors to be normalized ($\|\mathbf{e}_i\| = 1$). Solving the above differential equation for the covariance-based Hebb learning rule then gives us:

$$\mathbf{w}(t) = \sum_i c_i(0) e^{\frac{\lambda_i t}{\tau_w}} \mathbf{e}_i$$

where $\lambda_i$ is the eigenvalue of $\mathbf{C}$ corresponding to eigenvector $\mathbf{e}_i$. As $t \to \infty$, the eigenvector with the largest eigenvalue $\lambda_1$ dominates, giving

$$\lim_{t \to \infty} \mathbf{w}(t) \propto \mathbf{e}_1$$

as long as $\mathbf{w}(0)$ is not perpendicular to $\mathbf{e}_1$ (such that $c_1(0) = 0$). Thus, this learning rule leads to post-synaptic activity $v$ proportional to the projection of pre-synaptic input $\mathbf{u}$ onto the direction of maximum variance of the input patterns observed during learning, i.e. the principal eigenvector of the covariance matrix $\mathbf{C}$:

$$v \propto \mathbf{e}_1^\mathrm{T}\mathbf{u}$$

A similar analysis holds for the basic correlation-based Hebb rule with no thresholds, with the exact same result whenever the inputs have mean 0 such that $\mathbf{Q} = \mathbf{C}$ (see D&A fig. 8.4 for the difference whenever $\mathbf{Q} \neq \mathbf{C}$).

This analysis only holds, however, if the weights are allowed to change unboundedly, which does not occur with real synapses. Not only can they not grow unboundedly, but they cannot switch

from excitatory to inhibitory. Constraining all synapses to be excitatory, we should thus impose a lower bound at 0 along with a reasonable upper bound. In this case, the above theoretical result will still hold as long as the initial condition $\mathbf{w}(0)$ is far enough away from any of the boundaries such that there is enough time for $\mathbf{e}_1$ to dominate the dynamics.

In fact, the consideration of bounding synaptic weight changes brings up two general problems with the basic Hebb rule:

1. Without any bounds on the size of the $w_i$, the Hebb rule is unstable:

$$\tau_w \frac{\mathrm{d}\|\mathbf{w}\|}{\mathrm{d}t} = 2\mathbf{w}^{\mathrm{T}} \tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = 2\mathbf{w}^{\mathrm{T}} \mathbf{C}\mathbf{w}$$

   which is necessarily always positive since $\mathbf{C}$ is a covariance matrix and therefore positive semi-definite. In other words, as long as the learning rule is in effect the weights will continue increasing. In a network setting, this will quickly lead to runaway excitation.

2. Since all synaptic weights $w_i$ are allowed to grow unboundedly simultaneously, there is no competition between them. In light of the above the result, in a population of neurons receiving the same feed-forward input, Hebbian learning on the feed-forward weights will lead to a highly redundant representation whereby each neuron is activated in exactly the same way by each input (i.e. proportional to its projection onto the principal eigenvector of $\mathbf{C}$). This greatly limits the power of Hebbian learning in a network.

We now turn to two extensions of the basic Hebb rule that fix these issues.

## 4.2 BCM rule

The BCM rule (Bienenstock, Cooper & Munro, 1982) addresses the stability and competition issues by introducing a dynamic "sliding threshold":

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = v\mathbf{u}(v - \theta_v)$$
$$\tau_\theta \frac{\mathrm{d}\theta_v}{\mathrm{d}t} = v^2 - \theta_v$$

where $\tau_\theta < \tau_w$. Since an increase in $\|\mathbf{w}\|$ leads to an increase in $v$, the sliding threshold enforces stability by quickly and strongly increasing the threshold for LTP to prevent the weights from increasing further. Furthermore, a large increase in $w_i$ can make this happen in the absence of growth in the other $w_{j\neq i}$, thus preventing them from growing by pushing up the threshold, effectively implementing competition between weights.

## 4.3 Synaptic Normalization

The BCM rule effectively implements weight stability and competittion by using the post-synaptic activity $v$ as a proxy for the size of the weights. Alternatively, we could directly address the weight strengths by directly constraining the $L_p$ norm $\|\mathbf{w}\|_p = \sum_i w_i^p$. This is called *synaptic normalization*. Constraining the $L_1$ norm leads to *subtractive normalization*, whereas constraining the $L_2$ norm leads to *multiplicative normalization*.

### 4.3.1 Subtractive Normalization

The learning rule that constrains the $L_1$ norm of $\mathbf{w}$ is given by

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = v(\mathbf{u} - \bar{u}\mathbf{1})$$

where $\mathbf{1}$ is a vector of ones and

$$\bar{u} = \frac{1}{N_u}\sum_{i=1}^{K} u_k = \frac{\mathbf{1}^{\mathrm{T}}\mathbf{u}}{N_u} = \frac{\|\mathbf{u}\|_1}{N_u}$$

is a scalar, where $N_u$ is the number of pre-synaptic inputs. It is easy to verify that this rule constrains the total sum of synaptic weights:

$$\frac{\mathrm{d}\|\mathbf{w}\|_1}{\mathrm{d}t} = \sum_i \frac{\mathrm{d}w_i}{\mathrm{d}t} = v\sum_i u_i - vN_u\bar{u} = 0$$

Since the same quantity $v\bar{u}$ is being subtracted from the derivatives of each weight $\frac{\mathrm{d}w_i}{\mathrm{d}t}$, this synaptic normalization rule is termed subtractive.

As above, we consider this rule in the expectation over inputs, plugging in the equilibrium value for $v$:

$$\tau_w \left\langle \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} \right\rangle = \langle v\mathbf{u}\rangle - \langle \bar{u}v\rangle\mathbf{1} = \langle \mathbf{u}\mathbf{u}^{\mathrm{T}}\rangle\mathbf{w} - \frac{1}{N_u}\mathbf{1}^{\mathrm{T}}\langle \mathbf{u}\mathbf{u}^{\mathrm{T}}\rangle\mathbf{w}\mathbf{1} = \mathbf{Q}\mathbf{w} - \frac{\mathbf{1}^{\mathrm{T}}\mathbf{Q}\mathbf{w}}{N_u}\mathbf{1}$$

Taking the eigenvalue expansion of $\mathbf{Q}$ and recalling that its eigenvalues are orthogonal (because $\mathbf{Q}$ is symmetric) such that $\mathbf{w}(t) = \sum_i c_i(t)\mathbf{e}_i, c_i(t) = \mathbf{e}_i^{\mathrm{T}}\mathbf{w}$, we have:

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \sum_{i=1}^{N_u} \lambda_i c_i(t)\mathbf{e}_i - \frac{\lambda_i c_i(t)\mathbf{1}^{\mathrm{T}}\mathbf{e}_i}{N_u}\mathbf{1}$$

Recalling that the orthogonality of the eigenvectors of $\mathbf{Q}$ implies $\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j = \delta_{ij}$, we note that we have a differential equation for each $c_j(t)$:

$$\tau_w \frac{\mathrm{d}c_j}{\mathrm{d}t} = \tau_w \mathbf{e}_j^{\mathrm{T}}\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \lambda_j c_j(t) - \frac{\sum_{i=1}^{N_u}\lambda_i c_i(t)\mathbf{1}^{\mathrm{T}}\mathbf{e}_i}{N_u}\mathbf{e}_j^{\mathrm{T}}\mathbf{1} = \lambda_j c_j(t) - \frac{\sum_{i=1}^{N_u}\lambda_i c_i(t)\mathbf{1}^{\mathrm{T}}\mathbf{e}_i}{\sqrt{N_u}}\cos\theta_j$$

where $\theta_j$ is the angle between the vectors $\mathbf{e}_j$ and $\mathbf{1}$. We thus see that the subtractive normalization only operates on directions of growth of $\mathbf{w}$ close to the identity line $\mathbf{1}$, i.e. directions $\mathbf{e}_j$ in which all the weights grow at about the same rate. Directions of growth $\mathbf{e}_j$ perpendicular to $\mathbf{1}$ are left unaffected, since the normalization term in the derivative of the corresponding coefficient $c_j$ will be 0, leaving only standard Hebbian dynamics $\tau_w\frac{\mathrm{d}c_j}{\mathrm{d}t} = \lambda_j c_j(t)$ (exponential growth with rate equal to the corresponding eigenvalue). Consider the case where $\mathbf{e}_j \propto \mathbf{1}$, such that $\cos\theta_i = \delta_{ij} \Leftrightarrow \mathbf{e}_i^{\mathrm{T}}\mathbf{1} = \delta_{ij}\sqrt{N_u}$. We then have:

$$\begin{aligned}
\tau_w\frac{\mathrm{d}c_i}{\mathrm{d}t} &= \lambda_i c_i(t) - \lambda_j c_j(t)\delta_{ij} = (1-\delta_{ij})\lambda_i c_i(t) \\
&\Rightarrow c_i(t) = c_i(0)e^{\frac{(1-\delta_{ij})\lambda_i t}{\tau_w}} \\
&\Rightarrow \mathbf{w}(t) = c_j(0)\mathbf{e}_j + \sum_{i\neq j} c_i(0)e^{\frac{\lambda_i t}{\tau_w}}\mathbf{e}_i
\end{aligned}$$

If $\mathbf{e}_j \propto \mathbf{1}$ happens to be the principal eigenvector of $\mathbf{Q}$, then in the limit $t \to \infty$ $\mathbf{w}$ will grow in the direction of the eigenvector with the second highest eigenvalue. Otherwise, the long run limit is unaffected by subtractive normalization. The above analysis is easily generalized to the covariance-based Hebb rule by simply incorporating a pre- or post- synaptic activity threshold.

We can use this rule to model ocular dominance in a post-synaptic neuron. Consider two inputs $\mathbf{u} = \begin{bmatrix} u_R & u_L \end{bmatrix}^{\mathrm{T}}$ coming from each eye. Assuming the statistics to each eye are the same, we have:

$$\mathbf{Q} = \begin{bmatrix} \langle u_R u_R\rangle & \langle u_R u_L\rangle \\ \langle u_R u_L\rangle & \langle u_L u_L\rangle \end{bmatrix} = \begin{bmatrix} q_v & q_c \\ q_c & q_v \end{bmatrix}$$

which has two eigenvectors

$$\begin{aligned}
\mathbf{e}_1 &= \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \lambda_1 = q_v + q_c \\
\mathbf{e}_2 &= \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}, \lambda_2 = q_v - q_c
\end{aligned}$$

Since there is likely to be some covariance between the inputs to the two eyes, $q_c > 0 \Rightarrow \lambda_1 > \lambda_2$ and $\mathbf{e}_1$ is the principal eigenvector. Thus, without any normalization, Hebbian learning of the feed-forward input weights will lead to strengthened input connections from both eyes. But since

$\mathbf{e}_1 \propto \mathbf{1}$, subtractive normalization can force the weight vector to grow in the direction of $\mathbf{e}_2$ instead. Imposing a lower bound of 0 on the weights, this will lead to a 0 weight for one of the eyes and a large weight for the other (equal to the sum over the initial weights, since the subtractive normalization enforces $\frac{\mathrm{d}\|\mathbf{w}\|_1}{\mathrm{d}t} = 0$), thus producing ocular dominance in the post-synaptic activity.

Now for some caveats. From a biological perspective, the subtractive normalization rule is difficult to reconcile with known synaptic mechanics since it requires normalization by a global signal $\bar{u}$. In other words, the weight dynamics at synapse $i$ requires knowledge of the inputs at all other synapses. It is unclear how this could happen biologically. Another characteristic of subtractive normalization is that the competition between weights is very strong, since the global subtractive term is relatively stronger for weights with smaller derivatives. Indeed, without a lower bound on the weights, this rule will lead to driving weights below 0. With a bound at 0, it will tend to produce a solution with 1 big positive weight and all others set to 0.

### 4.3.2 Multiplicative Normalization

Also called the Oja Rule (Oja, 1982), the synaptic learning rule that constrains the $L_2$ norm of $\mathbf{w}$ is given by

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = v\mathbf{u} - \alpha v^2 \mathbf{w}$$

where $\alpha$ is a positive constant that bounds the $L_2$ norm of $\mathbf{w}$:

$$\frac{\mathrm{d}\|\mathbf{w}\|_2}{\mathrm{d}t} = 2\mathbf{w}^{\mathrm{T}}(v\mathbf{u} - \alpha v^2 \mathbf{w}) = 2v^2(1 - \alpha\|\mathbf{w}\|_2)$$

which converges to $\|\mathbf{w}\|_2 = \frac{1}{\alpha}$. Since the normalization term $\alpha v^2 w_i$ is proportional to each weight, we call this type of synaptic normalization multiplicative.

The Oja rule leaves our result for covariance-based synaptic learning rules intact, with the addition of a guarantee of convergence. Specifically, one can easily verify that this rule will lead to $\mathbf{w}(t) \to \mathbf{e}_1/\sqrt{\alpha}$ as $t \to \infty$. The Oja rule is also more biologically feasible than subtractive normalization since it only requires local signals for each synaptic weight change.

## 4.4 Spike-Timing Dependent Plasticity (STDP)

Another classical experimental finding is *spike-timing-dependent plasticity (STDP)* (Markram et al, 1997; Bi & Poo, 1998), whereby synaptic plasticity is only induced when action potentials in the pre- and post- synaptic cells occur within $\sim 50$ms of each other, the magnitude of the plasticity decaying with increasing latency. If the post-synaptic spike occurs after the pre-synaptic spike, LTP occurs; if, on the other hand, the post-synaptic spike precedes the pre-synaptic spike, then LTD occurs. The classical picture can be seen in D&A fig. 8.2. This can be easily implemented in spiking networks. In a continuous activity network like those we have been considering in this section, we can approximate it using a function $H(\tau)$ that gives the weight change when $t_{post} - t_{pre} = \tau$, $t_{post}, t_{pre}$ being adjacent post- and pre- synaptic spike times (e.g. the line in fig. 8.2B):

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \int_0^\infty \mathrm{d}\tau H(\tau) v(t)\mathbf{u}(t-\tau) + H(-\tau)v(t-\tau)\mathbf{u}(t)$$

where $sign(H(\tau)) = sign(\tau)$ such that the first and second terms inside the integral induce LTP and LTD respectively.

Like the basic Hebb rule, the STDP learning rule is unstable. It does, however, implement competition between weights: an increase in $w_i$ makes it easier for an increase in $u_i$ to lead to higher $v$ regardless of the state of the other inputs $u_{j\neq i}$, thus possibly increasing $v(t-\tau)u_{j\neq i}(t)$ and inducing LTD at those synapses. This tends to lead to a highly bimodal distribution of feed-forward weights.

Interestingly, the STDP rule can lead to invariant responses. We can approximately solve the above differential equation by ignoring the changes in $\mathbf{v}$ over time caused by the changes in $\mathbf{w}$:

$$\mathbf{w} = \frac{1}{\tau_w}\int_0^T \mathrm{d}t v(t) \int_{-\infty}^\infty \mathrm{d}\tau H(\tau)\mathbf{u}(t-\tau)$$

where we have also assumed $\mathbf{w}(0) = 0$ and ignored small contributions from the end points of the integral. In this approximation, our final learned $\mathbf{w}$ depends on the temporal correlation between

the post-synaptic activity $v(t)$ and the pre-synaptic activity $\mathbf{u}(t)$ temporally filtered by the STDP kernel $H(\tau)$. Consider now the scenario of $\mathbf{u}(t)$ arising from an object moving across the visual field. If the filter $H(\tau)$ filters the resulting sequence of inputs over the amount of time the object is present, then it will strengthen the synapses from all pre-synaptic cells responding to the object while it moves, regardless of its position. In the long run, the resulting weights will thus lead to post-synaptic responses independent of the position of the object, producing position-invariant responses to the object such as those seen in inferotemporal cortex (IT).

STDP can also produce predictive coding responses in a recurrent network with fixed feedforward weights. Consider a set of post-synaptic neurons with equally spaced homogenous tuning curves (e.g. for orientation) and an input stimulus that traverses the stimulus space in the same direction on each presentation (e.g. a clockwise rotating bar). As the stimulus is repeated, the tuning curves will then gradually shift in the opposite direction, since each neuron's recurrent input from neurons selective for the previous stimulus state will be strengthened. On each subsequent presentation, then, a given post-synaptic neuron is more and more likely of firing earlier and earlier. In the long run, this will produce predictive responses anticipating subsequent input according to the input stimulus it was trained on. Such behavior is observed in hippocampal place cells (D&A pgs 312-3).

## 4.5 Plasticity in a Network

As briefly mentioned above, our analysis of the basic Hebb rule implies that if we impose it on a feed-forward network of neurons with no recurrent connections, they will all learn the same input weights, aligned to the principal eigenvector of $\mathbf{Q}$ (or $\mathbf{C}$). This results in a completely redundant representation of the input where all post-synaptic neurons have exactly the same response to every input. By adding recurrent connections to the network, we can hope to avoid this.

We consider a linear recurrent network, of the form

$$\tau \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\mathbf{v} + \mathbf{W}\mathbf{u} + \mathbf{M}\mathbf{v}$$

where $\mathbf{W}$ are the feed-forward connection weights and $\mathbf{M}$ are the recurrent connection weights. Again assuming $\tau << \tau_w$, we will take the network activity at its stable equilibrium [10]:

$$\mathbf{v} = \mathbf{W}\mathbf{u} + \mathbf{M}\mathbf{v} \Leftrightarrow \mathbf{v} = \mathbf{K}\mathbf{W}\mathbf{u}$$

where $\mathbf{K} = (\mathbf{I} - \mathbf{M})^{-1}$. The basic Hebb rule is then:

$$\tau_w \frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t} = \langle \mathbf{v}\mathbf{u}^{\mathrm{T}} \rangle = \mathbf{K}\mathbf{W}\mathbf{Q}$$

Under certain conditions on $\mathbf{K}$, we can use subtractive normalization to generate ocular dominance bands in a population of neurons arranged in 1D space. Consider visual input from each eye $\mathbf{u} = \begin{bmatrix} u_R & u_L \end{bmatrix}^{\mathrm{T}}$ as above, with $\mathbf{Q}$ again as before. In this case, $\mathbf{W} = \begin{bmatrix} \mathbf{w}_R & \mathbf{w}_L \end{bmatrix}$ is an $N \times 2$ matrix, where $N$ is the number of neurons in the population. Using Hebbian learning with subtractive normalization for each set of feed-forward weights $\begin{bmatrix} w_{Ri} & w_{Li} \end{bmatrix}$ to each neuron (i.e. each row of $\mathbf{W}$) ensures that $\mathbf{w}_+ = \mathbf{w}_R + \mathbf{w}_L$ be constant over time. Thus, we can write

$$\tau_w \frac{\mathrm{d}\mathbf{w}_+}{\mathrm{d}t} = \tau_w \frac{\mathrm{d}\mathbf{w}_R}{\mathrm{d}t} + \tau_w \frac{\mathrm{d}\mathbf{w}_L}{\mathrm{d}t} = 0$$
$$\Rightarrow \tau_w \frac{\mathrm{d}\mathbf{w}_-}{\mathrm{d}t} = \tau_w \frac{\mathrm{d}\mathbf{w}_R}{\mathrm{d}t} - \tau_w \frac{\mathrm{d}\mathbf{w}_L}{\mathrm{d}t} = (q_v - q_c)\mathbf{K}\mathbf{w}_- \neq 0$$

as long as the weights are changing. Components of $\mathbf{w}_- = \mathbf{w}_R - \mathbf{w}_L$ that are highly positive indicate post-synaptic activity $v_i$ dominated by right eye input, whereas those that are highly negative reflect dominance by left eye input. If $\mathbf{K}$ is translation invariant - $K_{ij} = f(|i - j|)$ where the indexes $i, j$ designate each post-synaptic neuron's position in 1D space as well as their position

[10] Rewriting the dynamics as

$$\tau \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = (\mathbf{M} - \mathbf{I})\mathbf{v} + \mathbf{W}\mathbf{u}$$

we see that the recurrent network will have stable dynamics as long as the largest eigenvalue of $\mathbf{M} - \mathbf{I}$ is less than 0, i.e. the largest eigenvalue of $\mathbf{M}$ is less than 1.

in the matrix - then one can show that the principle eigenvectors of $\mathbf{K}$ lead to ocular dominance bands (see D&A pgs. 303-304).

An alternative approach to ensure Hebbian learning doesn't lead to redundant representation is to impose a non-linearity in the network that induces competition in post-synaptic activity. Purely linear recurrent interactions induce very weak competition, limiting the amount of differentiation achievable through Hebbian learning. So we introduce divisive normalization in post-synaptic activity:

$$v_i = \sum_j M_{ij} z_j$$
$$z_i = \frac{(\mathbf{w}_i^{\mathrm{T}} \mathbf{u})^\alpha}{\sum_j (\mathbf{w}_j^{\mathrm{T}} \mathbf{u})^\alpha}$$

where $\alpha$ controls the strength of competition. For large $\alpha$, for example, only $z_i$ with largest $\mathbf{w}_i^{\mathrm{T}} \mathbf{u}$ survive, with all others reduced to near 0. While the latter equation implements competition for feedforward input between all post-synaptic neurons, the first equation allows for cooperation between nearby neurons when the recurrent connections $\mathbf{M}$ are excitatory and local. Running Hebbian learning on the feedforward weights $\mathbf{w}_i$ in this setting is called *competitive Hebbian learning*. The nonlinear competition allows for strong differentiation of the post-synaptic neurons, which depend on higher order statistics beyond covariances (so in this case we cannot analyze the weight dynamics purely in terms of the eigenvectors of the correlation/covariance matrix). Indeed, the basic Hebb rule in this setting can produce ocular dominance bands in a population without the need for subtractive normalization.

By abstracting away from any physical grounding, competitive Hebbian learning can allow the formation of highly structured cortical maps. In such models, called *competitive feature-based models*, we assume the inputs $u_i$ take on the values of different parameters of the stimulus (e.g. ocularity, orientation, location), such that $N_u$ is the number of parameters used to characterize the stimulus and $W_{ij}$ directly represents the selectivity of neuron $i$ for parameter $j$. We can then accordingly modify our post-synaptic activity variable, e.g. by assuming homogenous Gaussian tuning curves for each parameter:

$$z_i = \frac{\exp\left[-\sum_j \frac{(u_j - W_{ij})^2}{2\sigma_j^2}\right]}{\sum_n \exp\left[-\sum_j \frac{(u_j - W_{nj})^2}{2\sigma_j^2}\right]}$$

The cooperation can then be introduced either via recurrent connections (self-organizing map) or via an extra cooperative term in the learning rule (elastic net). In the *self-organizing map* model, we assume the above equation for $v_i$, with $\mathbf{M}$ such that all recurrent connections are excitatory and local, to generate similar selectivities in nearby neurons. We then modify the basic Hebb rule to push a neuron's selectivity towards those inputs that excite it most:

$$\tau_w \frac{\mathrm{d}W_{ij}}{\mathrm{d}t} = \langle v_i (u_j - W_{ij}) \rangle$$

This is called the *feature-based learning rule*. The alternative is the *elastic net* model, which assumes $v_i = z_i$ and instead introduces an extra term in the learning rule to encourage similar selectivities between nearby neurons:

$$\tau_w \frac{\mathrm{d}W_{ij}}{\mathrm{d}t} = \langle v_i (u_j - W_{ij}) \rangle + \beta \sum_{n \in \text{neighborhood of } i} W_{nj} - W_{ij}$$

This is called the *elastic net rule*. Using ocularity, orientation, and location as stimulus parameters, these two models can produce orientation and ocular dominance cortical maps akin to those found in primates (i.e. with ocular dominance bands and iso-orientation countours with pinwheels; D&A fig. 8.10).

Note that the above models of plasticity in a network assume fixed non-plastic recurrent weights $\mathbf{M}$. Instead of making this strong assumption (likely to be false), we could instead apply synaptic learning rules to learn both the feedforward and recurrent weights simultaneously. By using a Hebbian rule for the feedforward weights and an anti-Hebbian rule for the recurrent weights, we

can then hope to decorrelate the post-synaptic activity to avoid redundant representation:

$$\tau_w \frac{\mathrm{d}W_{ij}}{\mathrm{d}t} = \langle v_i u_j \rangle - \alpha \langle v_i^2 \rangle W_{ij}$$
$$\tau_M \frac{\mathrm{d}M_{ij}}{\mathrm{d}t} = -\langle v_i v_j \rangle + \beta M_{ij}$$

with $M_{ii}$ set constantly to 0. By incorporating multiplicative normalization for the feedforward weights (i.e. Oja rule) and picking a suitable $\tau_M$ and $\beta$, this rule results in individual feedforward weights $\mathbf{w}_i$ (i.e. rows of $\mathbf{W}$) aligned to different eigenvectors of $\mathbf{Q} = \langle \mathbf{u}\mathbf{u}^{\mathrm{T}} \rangle$, and $M_{ij} = 0$. Indeed, anti-Hebbian plasticity is thought to be the predominant form of plasticity at the synapses of parallel fibres onto Purkinje cells in the cerebellum.

Alternatively, we could directly address the issue of redundant representation in a network by deriving a learning rule for $\mathbf{M}$ that sets correlations in post-synaptic activity to 0, i.e.

$$\langle \mathbf{v}\mathbf{v}^{\mathrm{T}} \rangle = c\mathbf{I}$$

By substituting in the equilibrium value of $\mathbf{v}$, we have

$$\begin{aligned} \langle \mathbf{v}\mathbf{v}^{\mathrm{T}} \rangle &= \mathbf{K}\mathbf{W}\langle \mathbf{u}\mathbf{v}^{\mathrm{T}} \rangle = c\mathbf{I} \\ \Leftrightarrow c\mathbf{K}^{-1} &= \mathbf{W}\langle \mathbf{u}\mathbf{v}^{\mathrm{T}} \rangle \\ \Leftrightarrow \mathbf{M} &= \mathbf{I} - c^{-1}\mathbf{W}\langle \mathbf{u}\mathbf{v}^{\mathrm{T}} \rangle \end{aligned}$$

This gives us the anti-Hebbian *Goodall rule*:

$$\tau_M \frac{\mathrm{d}\mathbf{M}}{\mathrm{d}t} = \mathbf{I} - \mathbf{W}\langle \mathbf{u}\mathbf{v}^{\mathrm{T}} \rangle - \mathbf{M}$$

which, if it converges, will converge to the desired matrix derived above. In addition to decorrelating post-synaptic activity, it also equates the individual variances. The Goodall rule, however, is non-local because of the $\mathbf{Wu}$ term, which implies the dynamics of any *recurrent* synapse weight $M_{ij}$ will depend on the activity at all the *feedforward* synapses on neuron $i$ (although this could be easily addressed by assuming the feedforward mapping is the identity, i.e. $\mathbf{W} = \mathbf{I}$). Our above result also requires the diagonal of $\mathbf{M}$ to be non-zero, so autapses are implied. These two characteristics make this rule somewhat biologically implausible. Computationally, it is also limited by its purely linear dynamics, which limit it to only removing redundancies in second-order statistics.

## 4.6 Plasticity for Supervised Learning

So far, we have only considered unsupervised learning rules, where the network is asked to learn some useful set of weights given a set of training inputs. Although somewhat less relevant biologically, we might also ask how Hebbian learning performs in a supervised setting, where we want the weights to be modified such that a given target output $\mathbf{v}$ is achieved in response to a corresponding input $\mathbf{u}$. It turns out that an interplay of Hebbian and anti-Hebbian learning is again crucial. We first analyze the case of applying Hebbian learning to see its limitations, which we then address with anti-Hebbian learning.

We first consider the problem of binary classification, using the *perceptron* binary classifier model:

$$v = \begin{cases} +1 & \text{if } \mathbf{w}^{\mathrm{T}}\mathbf{u} - \gamma \geq 0 \\ -1 & \text{else} \end{cases}$$

where the components of $\mathbf{u}$ are also either 1 or $-1$. We analyze the simplified case of $\gamma = 0$, using the basic Hebb rule with multiplicative normalization:

$$\tau_w \frac{\mathrm{d}w}{\mathrm{d}t} = \langle v\mathbf{u} \rangle - \alpha\mathbf{w} = \frac{1}{D}\sum_{i=1}^{D} v^{(i)}\mathbf{u}^{(i)} - \alpha\mathbf{w}$$

where $D$ is the number of observed data points. Note that the post-synaptic activity is indexed as well, since in the supervised setting we observe input-output pairs $(v^{(i)}, \mathbf{u}^{(i)})$ rather than just

inputs. We assume a random set of such pairs and ask how well the perceptron has learned the trained associations on convergence of the Hebbian learning rule to

$$\mathbf{w} = \frac{1}{\alpha D}\sum_{i=1}^{D} v^{(i)}\mathbf{u}^{(i)}$$

Setting $\alpha = \frac{K}{D}$, where $K$ is the number of input components, we have that for an arbitrary input $\mathbf{u}^{(j)}$ from the training data set, the perceptron will output

$$v\left(\mathbf{u}^{(j)}\right) = \mathbf{w}^{\mathrm{T}}\mathbf{u}^{(j)} = \frac{1}{K}\sum_{i=1}^{D} v^{(i)}{\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)} = \frac{v^{(j)}{\mathbf{u}^{(j)}}^{\mathrm{T}}\mathbf{u}^{(j)}}{K} + \frac{1}{K}\sum_{i\neq j}^{D} v^{(i)}{\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)} = v^{(j)} + \eta^{(j)}$$

which looks like a noisy version of the desired output. Indeed, in the limit of large $D$, the Central Limit Theorem tells us that

$$\lim_{D\to\infty}\frac{K}{D-1}\eta^{(j)} = \lim_{D\to\infty}\frac{1}{D-1}\sum_{i\neq j}^{D} v^{(i)}{\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{D-1}\right)$$

where $\mu, \sigma^2$ are the mean and variance of each of the $D-1$ terms in the sum[11]:

$$\mu = \langle v^{(i)}{\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)}\rangle = 0$$
$$\sigma^2 = \mathrm{Var}(v^{(i)}{\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)}) = K$$

Thus, for large $D$,

$$\eta^{(j)} \sim \mathcal{N}\left(0, \frac{D-1}{K}\right)$$

(since $\mathrm{Var}[\eta] = \left(\frac{D-1}{K}\right)^2 \mathrm{Var}\left[\frac{K}{D-1}\eta\right] = \left(\frac{D-1}{K}\right)^2 \frac{K}{D-1} = \frac{D-1}{K}$. In this limit, we can get a closed form expression for the probability that the perceptron correctly classifies an arbitrary input from the training set:

$$\begin{aligned} P\left(v(\mathbf{u}^{(j)}) = v^{(j)}\right) &= P(v^{(j)}=1)P(v(\mathbf{u}^{(j)})=1|v^{(j)}=1) + P(v^{(j)}=-1)P(v(\mathbf{u}^{(j)})=0|v^{(j)}=-1) \\ &= 0.5P(\eta^{(j)} > -1) + 0.5P(\eta^{(j)} < 1) \\ &= \mathrm{erf}\left(\sqrt{\frac{K}{D-1}}\right) \end{aligned}$$

[11] I found this to be a surprisingly non-trivial result, so I derive it here. Let $z$ be the number of positive terms in the sum implied by the dot product ${\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)}$. Since we have assumed each input component and output to be random and independent, $z \sim \mathrm{Binom}(K, 0.5)$. Recalling that $u_k \in \{+1, -1\}$, we can rewrite the dot product as ${\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)} = z - (K-z) = 2z - K$. We thus have:

$$\begin{aligned} \mu &= \langle v^{(i)}{\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)}\rangle \\ &= 2\langle v^{(i)}z\rangle - \langle (v^{(i)})\rangle K \\ &= 2\left(\sum_{a=0}^{K} P(v^{(i)}=+1)P(z=a)a + \sum_{a=-K}^{0} P(v^{(i)}=-1)P(z=-a)a\right) \\ &= \sum_{a=0}^{K} P(z=a)a - \sum_{a=0}^{K} P(z=a)a = 0 \\ \sigma^2 &= \langle (v^{(i)}{\mathbf{u}^{(i)}}^{\mathrm{T}}\mathbf{u}^{(j)})^2\rangle - \mu^2 \\ &= \langle {v^{(i)}}^2(2z-K)^2\rangle \\ &= 4\langle z^2\rangle - 4K\langle z\rangle + K^2 \quad \text{since } {v^{(i)}}^2 = 1 \text{ always} \\ &= 4(\mathrm{Var}[z] + (0.5K)^2) - 4K(0.5K) + K^2 \\ &= 4(0.25K + 0.25K^2) - 2K^2 + K^2 \\ &= K \end{aligned}$$

where erf$(\cdot)$ is the error function, or standard Gaussian cumulative distribution. Of course, this gives us a measure only of the *storage* capabilities of the perceptron following Hebbian learning with a large training data set. It is thus intuitive that the perceptron should be able to store trained associations when the dimensionality of the input is greater than the number of inputs ($K > D$), in which case it is easy to find a manifold that can linearly separate the trained inputs[12]. We have no guarantees for *generalization* when the input associations have some structure to be learned - in fact, the setting of $K > D$ is likely to lead to overfitting. Furthermore, this is a highly restrictive setting: we need a large $D$ for the above theoretical results (using the CLT) to hold, and an even larger $K$ to achieve good storage. One of the main reasons for this highly limited performance is that the Hebbian learning rule is sensitive only to correlation structure in the training data, with no regard to the actual responses of the perceptron. We might hope to do better by gauging the weight updates according to the direction of any errors the perceptron is making. This leads to the *perceptron learning rule*:

$$\mathbf{w} \to \mathbf{w} + \epsilon(v^{(i)} - v(\mathbf{u}^{(i)}))\mathbf{u}^{(i)}, \quad \gamma \to \gamma - \epsilon(v^{(i)} - v(\mathbf{u}^{(i)}))$$

which increases the weights whenever the perceptron misclassifies in the negative direction ($-1$ instead of $+1$, such that $v^{(i)} - v(\mathbf{u}^{(i)}) = 2 > 0$) and decreases them when it makes an error in the opposite direction (such that $v^{(i)} - v(\mathbf{u}^{(i)}) = -2 < 0$). If it correctly classifies the given input, the weights are left unchanged. Thus, this learning rule requires repeated exposure to all inputs, in which case it can be proved that it will converge to perfect classification whenever each class of inputs (i.e. $+1, -1$) are linearly separable. Note that if we expand the weight update rule by multiplying out the terms inside the parenthesis we get a Hebbian and an anti-Hebbian term.

We can get further insight into why we should include such anti-Hebbian learning by considering supervised learning compliment to classification: regression. Here, we want to approximate an arbitrary function $h(u)$ via a linear sum of *basis functions* $f_i(u)$:

$$h(u) \approx v(u) = \sum_i w_i f_i(u)$$

If the set of basis functions $\{f_i(\cdot)\}$ can represent a class of functions via a linear sum, we say that it is *complete* with respect to that class. If the set of corresponding weights $\{w_i\}$ is not unique for the given target function, then we say it is *overcomplete*. From a neural perspective, we might consider a population of $N$ neurons with tuning curves $f_i(u)$ responding to a stimulus $u$. A downstream post-synaptic neuron that linearly sums its pre-synaptic inputs can then represent any function of the stimulus in a class such that the $N$ tuning curves form a complete set of basis functions for it. More generally, for any arbitrary target function $h(u)$, we can hope that the post-synaptic activity $v(u)$ provides a good approximation of $h(u)$ if the synaptic weights $\mathbf{w}$ minimize the mean squared error over some set of $N$ observed input stimuli, i.e.

$$\begin{aligned}
0 &= \frac{\partial}{\partial \mathbf{w}} \frac{1}{2} \sum_{i=1}^{N} \left( h(u^{(i)}) - v(u^{(i)}) \right)^2 \\
&= \sum_{i=1}^{N} \left( h(u^{(i)}) - v(u^{(i)}) \right) \mathbf{f}(u^{(i)}) \\
\Leftrightarrow \mathbf{w} &= \left( \sum_{i=1}^{N} \mathbf{f}(u^{(i)}) \mathbf{f}(u^{(i)})^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} h(u^{(i)}) \mathbf{f}(u^{(i)}) \\
&= \left\langle \mathbf{f}(u)\mathbf{f}(u)^{\mathrm{T}} \right\rangle^{-1} \left\langle h(u)\mathbf{f}(u) \right\rangle
\end{aligned}$$

where $\mathbf{f}(u) = \begin{bmatrix} f_1(u) & f_2(u) & \dots & f_N(u) \end{bmatrix}^{\mathrm{T}}$. Can we learn such a set of weights with a Hebbian learning rule? Consider again the basic Hebb rule with multiplicative normalization, which in this case would coverge to:

$$\mathbf{w} = \frac{1}{\alpha} \langle h(u)\mathbf{f}(u) \rangle$$

In light of our above result, this set of synaptic weights will yield a good approximation of $h(u)$ if $\left\langle \mathbf{f}(u)\mathbf{f}(u)^{\mathrm{T}} \right\rangle = \alpha \mathbf{I}$, i.e. if the post-synaptic activity is decorrelated across the set of observed

[12] Note that the perceptron is only capable of replicating associations that are linearly separable (which implies a maximum of $2K$ different associations).

stimuli. Given a fixed set of stimuli, this is called a *tight frame* condition on the basis functions. Specifically, it implies that for any two observed stimuli $u^{(i)} \neq u^{(j)}$, $\mathbf{f}(u^{(i)})$ is orthogonal to $\mathbf{f}(u^{(j)})$. Thus, Hebbian learning is only effective under highly restrictive conditions. We see now that the issue stems from the fact that the Hebbian learning rule does not account for correlations in the responses to different inputs. Indeed, our analysis shows that it will only perform well if post-synaptic responses are completely decorrelated. In this light, it is easy to see how adding an anti-Hebbian term should alleviate this problem by decorrelating the post-synaptic responses. In fact, an anti-Hebbian term falls out of the derivation the *delta rule*, which is exactly equivalent to stochastic gradient ascent on the mean squared error:

$$\mathbf{w} \to \mathbf{w} + \epsilon \left( h(u^{(i)}) - v(u^{(i)}) \right) \mathbf{f}(u^{(i)})$$

This rule is directly analagous to the perceptron learning rule above, incorporating an anti-Hebbian term that makes it sensitive to the direction of the post-synaptic neuron's errors to enforce the appropriate weight changes.

Turning now to stochastic networks, we can derive an analagous weight update rule for density estimation with a Boltzmann machine. In this case, minimizing the KL divergence between the output distribution of a Boltzmann machine and some target conditional or joint distribution (equivalent to maximizing the likelihood of the observed training data) results in a very similar weight update rule that again consists of the difference between a Hebbian and anti-Hebbian term. It is thus called a *contrastive Hebb rule*. Due to the stochastic nature of the networks output, in this case the two terms are in fact implemented in separate phases, emphasizing the role of the anti-Hebbian learning in decorrelating the network output:

1. *Wake phase*: the Boltzmann machine is fed data, and the feed-forward and/or recurrent weights are updated with a Hebb rule. The network thus learns the correlation structure in the training data.

2. *Sleep phase*: the Boltzmann machine generates random samples in response to data inputs, and the weights are updated with an anti-Hebbian rule. The network thus modifies its weights to decorralte its stochastic output.

See D&A pgs 322-6 for more details.

# 5 Neural Coding

## 5.1 Information Theory

Consider a stimulus $S$ that can take on $M$ different values, and we want to transmit a messages about its identity at any given trial or point in time. The most straight-forward way of doing this is to use binary code, requiring messages of length $L$ given by $2^L = M \Leftrightarrow L = \log_2 M$ But note that in this code all stimuli are encoded with the same length message. This is a wasteful use of bits when you consider that some stimuli will be more likely than others - our messages about the stimulus could be much shorter on average if we used shorter strings to represent the values that occured the most frequently. It turns out that the optimal length for a message encoding $S = s$, is in fact given by the inverse of its probability[13]:

$$L(s) = \log_2 \frac{1}{p(s)}$$

Using this code, the average length of our messages will then be:

$$H[p] = -\sum_{s=1}^{M} p(s) \log_2 p(s)$$

This is called the *entropy*, and it gives the minimum average message length (i.e. the average message length when using the optimal code) with respect to sending messages about some signal $S$ with probability distribution $p(S = s) = p(s)$. Thus, the entropy is a functional of the probability

[13]For a nice intuitive demonstration of this, see http://colah.github.io/posts/2015-09-Visual-Information/

mass function $p(s)$. In the limit of 0 uncertainty where $p(S = s) = \delta(s - s')$, one can easily see that the entropy goes to 0: we don't need to even send a message about the stimulus if everyone knows that it is always going to be $S = s'$. On the other hand, if the stimuli are uniformly distributed $p(S = s) = \frac{1}{M}$, the entropy reaches its maximum, which can be derived via Jensen's inequality:

$$\begin{aligned} H[p] &= \sum_{s=1}^{M} p(s) \log_2 \frac{1}{p(s)} \\ &\overset{\text{Jensen}}{\leq} \log_2 \sum_{s=1}^{M} \frac{p(s)}{p(s)} = \log_2 M = -\sum_{s=1}^{M} \frac{1}{M} \log_2 \frac{1}{M} \end{aligned}$$

Note that we can generalize all the above to any type of discrete code beyond binary values (i.e. bits). The only change required is to change the base of the logarithm to whatever unit of information is being used. The convention is to use *nats*, in which case we use the natural logarithm.

What if we don't use the optimal code? For example, what if we are mistaken about the true $p(s)$ and instead construct our code optimally with respect to some other distribution $q(s)$? Our average message length is then give by the *cross-entropy*:

$$H_q[p] = -\sum_s p(s) \log q(s)$$

which will be longer than the average message length using the optimal code. We can see this by simply taking the difference in message lengths under the two codes, given by the *Kullback-Leibler Divergence*:

$$\begin{aligned} \mathrm{KL}[p(s)||q(s)] &= H_q[p] - H[p] \\ &= -\sum_s p(s) \log q(s) + \sum_s p(s) \log p(s) \\ &= \sum_s p(s) \log \frac{p(s)}{q(s)} \\ \Rightarrow -\mathrm{KL}[p(s)||q(s)] &\overset{\text{Jensen}}{\leq} \log \sum_s p(s) \frac{q(s)}{p(s)} = 0 \\ \Leftrightarrow \mathrm{KL}[p(s)||q(s)] \geq 0 &\Leftrightarrow H_q[p] \geq H[p] \end{aligned}$$

where the inequality follows from Jensen's inequality.

Another thing we can ask is how to update our code in the face of new information about the stimulus. Consider, for example, that we observe the firing rate $r$ of a cell that responds to the stimulus $S$ according to the probability distribution $p(R = r|S)$. Using Baye's rule to get $p(S|R = r)$, we then update our code accordingly, using $p(s|R = r)$ instead of $p(s)$ for the distribution of the signal we are trying to communicate. Averaging over all possible responses $R$, we then have that our minimum average message length is given by the *conditional entropy*:

$$H[S|R] = \sum_r p(r) \left( -\sum_s p(s|r) \log p(s|r) \right) = -\sum_{r,s} p(s, r) \log p(s|r)$$

We can now use this to ask: how much does knowing $R = r$ improve our code? We can measure this by computing the resulting decrease in entropy (i.e. reduction in average message length), called the *mutual information*:

$$\begin{aligned} \mathrm{I}[S, R] &= H[S] - H[S|R] \\ &= -\sum_s p(s) \log p(s) + \sum_{r,s} p(s, r) \log p(s|r) \\ &= -\sum_{s,r} p(s, r) \log p(s) + \sum_{r,s} p(s, r) \log \frac{p(s, r)}{p(r)} \\ &= \sum_{s,r} p(s, r) \log \frac{p(s, r)}{p(s)p(r)} \end{aligned}$$

One can easily see that if the random variables $R, S$ are independent (such that $p(r,s) = p(r)p(s)$), their mutual information is 0, since in this case observing $R$ won't tell you anything about $S$ (so $p(s|r) = p(s)$ and therefore $H[S|R] = H[S]$). Note as well that the mutual information is symmteric (unlike the KL Divergence). Lastly, note that the mutual information is always positive: knowing more about the stimulus $S$ (i.e. conditioning $p(S) \to p(S|R)$) can only decrease your uncertainty about $S$ (i.e. decrease the entropy). One can see this most easily by reexpressing the mutual information as a Kullback-Leibler divergence:

$$\begin{aligned} &\mathrm{I}[S,R] = \mathrm{KL}[p(s,r)||p(s)p(r)] \geq 0 \\ \Leftrightarrow\ &H[S|R] \leq H[S] \end{aligned}$$

Suppose we apply a series of transformations to the stimulus $S$, giving us a Markov chain of the form $S \to R_1 \to R_2$ such that $R_2 \perp\!\!\!\perp S|R_1$. The *data processing inequality* then tells us that the information about $S$ contained in $R_2$ cannot be more than that contained in $R_1$: information can never increase. This follows from the fact that conditioning only decreases entropy, such that

$$\begin{aligned} &H[S|R_2] \geq H[S|R_1, R_2] = H[S|R_1] \\ \Leftrightarrow\ &H[S|R_2] \geq H[S|R_1] \Rightarrow \mathrm{I}[S,R_2] \leq \mathrm{I}[S,R_1] \end{aligned}$$

where the equality in the first line follows from the Markov independence structure. In more detail, we can prove this by again using the fact that the mutual information is always positive:

$$\begin{aligned} \forall r_2\ 0 &\leq \mathrm{I}[S,R_1|R_2 = r_2] \\ &= \sum_{s,r_1} p(s,r_1|r_2) \log \frac{p(s,r_1|r_2)}{p(s|r_2)p(r_1|r_2)} \\ &= \sum_{s,r_1} p(s,r_1|r_2) \log \frac{p(s|r_1,r_2)}{p(s|r_2)} \\ &= \sum_{s,r_1} p(s,r_1|r_2) \log p(s|r_1,r_2) - \sum_{s} p(s|r_2) \log p(s|r_2) \\ \Rightarrow 0 &\leq \underbrace{\sum_{s,r_1,r_2} p(r_2)p(s,r_1|r_2) \log p(s|r_1,r_2)}_{-H[S|R_1,R_2]} \underbrace{- \sum_{s,r_2} p(r_2)p(s|r_2) \log p(s|r_2)}_{H[S|R_2]} \\ &= \underbrace{\sum_{s,r_1,r_2} p(s,r_1,r_2) \log p(s|r_1)}_{-H[S|R_1]} + H[S|R_2] \\ &= -H[S|R_1] + H[S|R_2] \Leftrightarrow H[S|R_2] \geq H[S|R_1] \end{aligned}$$

where we used the fact that $p(s|r_1, r_2) = p(s|r_1)$ (by the Markov structure) to go from the fifth line to the sixth line, giving us $H[S|R_1, R_2] = H[S|R_1]$.

We can generalize the above notions to other probabilistic objects. Consider a stochastic process $\mathcal{S} = \{S_1, S_2, \ldots\}$. We define its *entropy rate* as

$$\begin{aligned} H[\mathcal{S}] &= \lim_{n\to\infty} \frac{H[S_1,\ldots,S_n]}{n} \\ &= \lim_{n\to\infty} \frac{H[S_n|S_{n-1},\ldots,S_1] + H[S_{n-1}|H_{n-2},\ldots,S_1] + \ldots + H[S_1]}{n} \\ &= \lim_{n\to\infty} H[S_n|H_{n-1},\ldots,S_1] \end{aligned}$$

where the last equality holds under the assumption of stationarity of the conditional distribution.

We can also generalize entropy to continuous random variables $s \in \mathbb{R}$ with probability density functions $p(s)$. In this case, however, the conventional definition of entropy breaks down, because you need an infinite number of bits to encode a real number. This becomes evident in directly deriving the entropy of $s$: to obtain probabilities from the probability density function $p(s)$, we must bin the possible values of $s \in \mathbb{R}$ into discrete equally sized bins $s_i$ and then take the limit of

infinitely small bin sizes $\Delta s \to 0$:

$$\begin{aligned} H[s] &= -\sum_i p(s_i)\Delta s \log p(s_i)\Delta s \\ &= -\sum_i p(s_i)\Delta s \log p(s_i) - \log \Delta s \\ &\overset{\Delta s \to 0}{\longrightarrow} -\int p(s)\log p(s)\mathrm{d}s + \infty \end{aligned}$$

Thus, for continuous random variables, we use the *differential entropy*

$$h(s) = -\int p(s)\log p(s)\mathrm{d}s$$

We can similarly define the conditional differential entropy, and in fact the KL divergence and mutual information need no modification since the $\infty$ terms cancel each other.

As illustrated above, the entropy of a random variable is a measure of our uncertainty about it (i.e. uncertainty $\propto$ optimal average message length). We can thus use it to derive the distribution (in a Bayesian sense) of a random variable whose true probability distribution we don't know. In the case that we know nothing about it, we have no idea what values are more likely than others, so we should assign it a uniform probability distribution - the probability distribution with highest entropy. But consider now the case where we know the mean $\mu$ of a random variable $X$ but nothing else. What probability distribution should we assign to it? Given that we know nothing else about $X$, we should assign the probability distribution with highest (differential) entropy, under the constraint that it have a mean set at $\mu$:

$$\begin{aligned} 0 &= \frac{\delta}{\delta p}\left[H[p] + \lambda_0\left(\int p(x)\mathrm{d}x - 1\right) + \lambda_1\left(\int xp(x)\mathrm{d}x - \mu\right)\right] \\ &= \frac{\delta}{\delta p}\left[-\int p(x)\log p(x) + \lambda_0\left(\int p(x)\mathrm{d}x - 1\right) + \lambda_1\left(\int xp(x)\mathrm{d}x - \mu\right)\right] \\ &= -\log p(x) - 1 + \lambda_0 + \lambda_1 x \\ \Leftrightarrow p(x) &\propto e^{\lambda_1 x} \end{aligned}$$

where $\lambda_0, \lambda_1$ are Lagrange multipliers to enforce the constraints the $p(x)$ be a probability distribution with mean $\mu$. In other words, if all we know is the mean of $X$, we should assume it is exponentially distributed, with that mean. This is the *maximum entropy distribution* of a random variable with fixed mean. Applying this result to the distribution of time intervals between events in a point process with a constant mean rate, we get that the *maximum entropy point process* with fixed constant mean rate is a homogenous Poisson process (with exponentialy distributed inter-event intervals). What if we also knew the variance $\sigma^2$ of $X$?

I'm not sure this is enough to prove this

$$\begin{aligned} 0 &= \frac{\delta}{\delta p}\left[H[p] + \lambda_0\left(\int p(x)\mathrm{d}x - 1\right) + \lambda_1\left(\int (x-\mu)^2 p(x)\mathrm{d}x - \sigma^2\right)\right] \\ &= -\log p(x) - 1 + \lambda_0 + \lambda_1(x-\mu)^2 \\ \Leftrightarrow p(x) &\propto e^{\lambda_1(x-\mu)^2} \end{aligned}$$

i.e. the maximum entropy distribution of a random variable with fixed mean and variance is Gaussian. We can verify this by considering an arbitrary probability density $q(x)$ with the same mean $\mu$ and variance $\sigma^2$ and calculating its KL divergence with the Gaussian $p(x)$:

$$\begin{aligned} \mathrm{KL}[q||p] &= \int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}x \\ &= -h(q) - \int q(x)\log p(x)\mathrm{d}x \\ &= -h(q) + \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\int q(x)(x-\mu)^2\mathrm{d}x \\ &= -h(q) + \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sigma^2 \end{aligned}$$

$$= -h(q) + \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}$$
$$= -h(q) + \frac{1}{2}\log 2\pi\sigma^2 e$$

Noting that

$$h(p) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\int p(x)(x-\mu)^2 \mathrm{d}x = \frac{1}{2}\log 2\pi\sigma^2 e$$

the positivity of the KL divergence gives us:

$$\mathrm{KL}[q||p] = -h(q) + h(p) \geq 0 \Leftrightarrow h(p) \geq h(q)$$

In other words, any probability distribution with the same mean and variance will have less entropy than the Gaussian.

Information theory can be useful for analyzing *channels*, whereby information about some source/stimulus $S$ is transmitted via a channel that outputs responses $R$ according to a probability distribution $p(R|S)$. This conditional probability distribution completely characterizes the channel, so we would like a way to measure how good the channel is without reference to the intrinsic properties of $S$ or $R$, such as their respective marginal distributions $p(S), p(R)$. The natural way to measure how useful the responses $R$ are for communicating information about $S$ is the mutual information $\mathrm{I}[S, R]$, but this indeed depends on $p(S)$ (and $p(R)$, but we can always calculate this using $p(r) = \sum_s p(s)p(r|s)$). We thus define the *capacity* of a channel to be

$$C_{R|S} = \sup_{p(S)} \mathrm{I}[S, R]$$

which gives the theoretical limit on the amount of information that can be transmitted through the channel. To use this channel optimally, then, you should ensure the distribution over inputs $p(S)$ saturates the channel capacity. This distribution can be found using an iterative EM-like algorithm called the Blahut-Arimoto algorithm.

However, usually we are constrained by a fixed input distribution $p(S)$. The problem then becomes finding an encoding $p(\tilde{S}|S)$ that maximizes $\mathrm{I}[S, R]$, where the encoded message $\tilde{S}$ is now the input to the channel. Given that, by the information-processing inequality,

$$\mathrm{I}[S, R] \leq \mathrm{I}[\tilde{S}, R] \leq C_{R|S}$$

one way to find a good $p(\tilde{S}|S)$ is to maximize $\mathrm{I}[\tilde{S}, R]$ so that our upper bound on $\mathrm{I}[S, R]$ saturates the channel capacity. We can do this by maximizing the marginal entropy over the channel outputs $H[R]$ with respect to $p(\tilde{S})$, since

$$\mathrm{I}[\tilde{S}, R] = H[R] - H[R|\tilde{S}]$$

Under no constraints on the moments of $R$, this translates to finding the encoding $p(\tilde{S}|S)$ such that the channel outputs $R$ are uniformly distributed (i.e. the maximum entropy distribution). One approach to this is called *histogram equalization*. Consider a deterministic encoding $\tilde{s} = f(s)$ and i.i.d. noise corrupted outputs $r = \tilde{s} + \eta$, where $\eta$ is a 0-mean random variable representing the noise. We want to find the deterministic function $f(s)$ that ensures $p(R = r) = \frac{1}{r_{max}}$ be uniform (between 0 and $r_{max}$, e.g. firing rate of a neuron). Since in this case $p(r) \approx p(\tilde{s})$, this implies deriving an encoding function $f(s)$ such that $p(\tilde{s}) = \frac{1}{r_{max}}$. Our first step is thus to express $p(\tilde{s})$ in terms of $f(s)$. Given $s, \tilde{s}$ such that

$$p(s)\mathrm{d}s = p(\tilde{s})\mathrm{d}\tilde{s}$$

we can express $p(\tilde{s})$ as

$$p(\tilde{s}) = p(s)\frac{\mathrm{d}s}{\mathrm{d}\tilde{s}} = p(s)\frac{1}{f'(s)}$$

Thus, our encoding function is given by

$$\frac{1}{r_{max}} = p(s)\frac{1}{f'(s)} \Leftrightarrow f(s) = r_{max}\int_{-\infty}^{s} p(s')\mathrm{d}s'$$

In other words, the encoding $s \to \tilde{s}$ consists of the maximum channel response $r_{max}$ scaled by the cumulative probability of the input $s$. Thus, the regions of $s$-space with highest probability density are diluted into very large regions of $\tilde{s}$-space, since the cumulative density changes the most over these regions. Regions of low probability density, on the other hand, are regions where the cumuluative density is relatively flat, so the encoding concentrates into smaller regions in $\tilde{s}$-space, thus spreading the probability equally around the encoded input $\tilde{s}$ space.

## 5.2 Fisher Information

Consider a neuron that fires $n$ spikes in response to a stimulus $\mathbf{s}$ according to a probability distribution $P(n|\mathbf{s})$. The question we are after is: *how much information does the neuron's spiking activity $n$ contain about $\mathbf{s}$?* A natural way of framing this is: *given $n$, how well can we estimate $\mathbf{s}$?*

We can get a reasonable answer by computing the posterior distribution $P(\mathbf{s}|n)$ in the infinite data limit, i.e. the limit of $N \to \infty$ in a data set $\{n_i\}_{i=1}^N$ of $N$ i.i.d. responses of the neuron to some stimulus $\mathbf{s}_0$:

$$\begin{aligned}\log P(\mathbf{s}|\{n_i\}) &= \log P(\{n_i\}|\mathbf{s}) + \log P(\mathbf{s}) - \log Z \\ &= \sum_{i=1}^{N} \log P(n_i|\mathbf{s}) + \log P(\mathbf{s}) - \log Z \\ \Rightarrow \lim_{N\to\infty} \log P(\mathbf{s}|\{n_i\}) &= N \left\langle \log P(n|\mathbf{s}) \right\rangle_{P(n|\mathbf{s}_0)} - \log Z'\end{aligned}$$

where in the last line we dropped all terms that don't scale with $N$ and used the law of large numbers to replace the infinite sum over $N$ data points with its mean, evaluated at the true stimulus value $\mathbf{s}_0$. Suppose now that, when estimating $\mathbf{s}$, we happen to know that it lies in the vicinity of $\mathbf{s}_0$. One might argue this is an unrealistically optimistic scenario to consider, but it will allow us to analyze our ability to estimate $\mathbf{s}$ at least for this idealistic case. In this case, we only need to evaluate the posterior at stimulus values near $\mathbf{s}_0$, so that we need only consider its Taylor expansion around $\mathbf{s}_0$ up to second order:

$$\begin{aligned}\log P(\mathbf{s}|\{n_i\}) \approx N \Bigg( &\langle \log P(n|\mathbf{s}_0) \rangle + (\mathbf{s}-\mathbf{s}_0)^{\mathrm{T}} \left\langle \frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s}) \right\rangle \\ &+ \frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^{\mathrm{T}} \left\langle \frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s}) \right\rangle (\mathbf{s}-\mathbf{s}_0) \Bigg) - \log Z' \\ = N(\mathbf{s}-\mathbf{s}_0)^{\mathrm{T}} &\left\langle \frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s}) \right\rangle - \frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^{\mathrm{T}} \left( -N \left\langle \frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s}) \right\rangle \right) (\mathbf{s}-\mathbf{s}_0) - \log Z''\end{aligned}$$

where we absorbed all terms constant w.r.t. $\mathbf{s}$ (namely, the first one) into the normalizer. We now note that

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s}) \right\rangle = \int \mathrm{d}n P(n|\mathbf{s}_0) \frac{1}{P(n|\mathbf{s}_0)} \frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} P(n|\mathbf{s}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} \int \mathrm{d}n P(n|\mathbf{s}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} 1 = 0$$

Exponentiating both sides we then arrive at:

$$P(\mathbf{s}|\{n_i\}) \approx \frac{1}{Z'} \exp\left[ -\frac{1}{2}(\mathbf{s}-\mathbf{s}_0)^{\mathrm{T}} \left( -N \left\langle \frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s}) \right\rangle \right) (\mathbf{s}-\mathbf{s}_0) \right] = \mathcal{N}\left( \mathbf{s} \middle| \mathbf{s}_0, \frac{1}{N}\mathbf{J}(\mathbf{s}_0)^{-1} \right)$$

where

$$\mathbf{J}(\mathbf{s}_0) = -\left\langle \frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s}) \right\rangle_{P(n|\mathbf{s}_0)}$$

is called the *Fisher information* matrix, giving us the rate at which the variance of the approximate Gaussian posterior in the large data limit decreases with increasing $N$. Recall, however, that this approximate posterior is only valid at values of $\mathbf{s}$ near the true stimulus $\mathbf{s}_0$.

We can get a more precise intuition for what the Fisher information means by realizing that it is in fact equal to the variance of the gradient of the log likelihood with respect to the stimulus

(also called the *score function*):

$$\begin{aligned}
\mathbf{J}(\mathbf{s}_0) &= -\left\langle \frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} \frac{1}{P(n|\mathbf{s})}\frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}P(n|\mathbf{s}) \right\rangle_{P(n|\mathbf{s}_0)} \\
&= -\left\langle \frac{1}{P(n|\mathbf{s}_0)}\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\bigg|_{\mathbf{s}_0} P(n|\mathbf{s}) \right\rangle_{P(n|\mathbf{s}_0)} + \left\langle \frac{1}{P(n|\mathbf{s}_0)^2}\left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} P(n|\mathbf{s})\right)\left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} P(n|\mathbf{s})\right) \right\rangle_{P(n|\mathbf{s}_0)} \\
&= -\int \mathrm{d}n \frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\bigg|_{\mathbf{s}_0} P(n|\mathbf{s}) + \left\langle \left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{s}}\bigg|_{\mathbf{s}_0} \log P(n|\mathbf{s})\right)^2 \right\rangle_{P(n|\mathbf{s}_0)} \\
&= \left\langle \left(\nabla \log P(n|\mathbf{s})\Big|_{\mathbf{s}_0}\right)\left(\nabla \log P(n|\mathbf{s})\Big|_{\mathbf{s}_0}\right)^{\mathrm{T}} \right\rangle_{P(n|\mathbf{s}_0)} \\
&= \mathrm{cov}_{P(n|\mathbf{s}_0)}\left[\nabla \log P(n|\mathbf{s})\Big|_{\mathbf{s}_0}\right]
\end{aligned}$$

where in the last line we recalled from above that $\langle \nabla \log P(n|\mathbf{s}) \rangle = 0$ such that the square of the derivative (the outer product of the gradient with itself) is in fact the (co)variance. For a one-dimensional stimulus, the Fisher information is thus easily interpreted as a measure of the sensitivity of the population response to changes in the stimulus (i.e. the squared derivative). Again, we note that the Fisher information is only a *local* measure of information, since the derivatives are evaluated at the stimulus value $\mathbf{s}_0$.

It also turns out that the Fisher information has a very specific statistical meaning even when $N$ is finite. Consider an estimator $\hat{s}(x)$ of the stimulus $s_0$, based on some data $x$ (e.g. a series of i.i.d. noisy measurements of $s_0$, a neural response, etc.), with bias given by

$$b(s_0) = \langle \hat{s}(x) \rangle_{P(x|s_0)} - s_0$$

We will see that the Fisher information gives a lower bound on the variance of this estimator. We first realize that we can express the square of the derivative of the bias in terms of the square of the expected derivative of the log likelihood:

$$\begin{aligned}
b'(s_0) &= \int \mathrm{d}x\ \hat{s}(x)\frac{\mathrm{d}}{\mathrm{d}s_0}P(x|s_0) - 1 \\
&= \int \mathrm{d}x\ \hat{s}(x)P(x|s_0)\frac{\mathrm{d}}{\mathrm{d}s_0}\log P(x|s_0) - 1 \\
&= \left\langle \hat{s}(x)\frac{\mathrm{d}}{\mathrm{d}s_0}\log P(x|s_0) \right\rangle_{P(x|s_0)} - 1 \\
\Rightarrow (b'(s_0)+1)^2 &= \left\langle \hat{s}(x)\frac{\mathrm{d}}{\mathrm{d}s_0}\log P(x|s_0) \right\rangle^2_{P(x|s_0)} \\
&= \left\langle (\hat{s}(x) - \langle \hat{s}(x) \rangle)\frac{\mathrm{d}}{\mathrm{d}s_0}\log P(x|s_0) \right\rangle^2_{P(x|s_0)}
\end{aligned}$$

where in the last line we simply realized that

$$-\langle \hat{s}(x) \rangle_{P(x|s_0)} \left\langle \frac{\mathrm{d}}{\mathrm{d}s_0}\log P(x|s_0) \right\rangle_{P(x|s_0)} = -\langle \hat{s}(x) \rangle_{P(x|s_0)} \left( \int \mathrm{d}x \frac{\mathrm{d}}{\mathrm{d}s_0}P(x|s_0) \right) = 0$$

so we are free to add in the term. By the *Cauchy-Schwarz inequality*[14], for any two random

[14]For any given inner product $\langle \cdot, \cdot \rangle$, it is always the case that $\langle u, v \rangle \leq \langle u, u \rangle \langle v, v \rangle$. For the above result, called the *covariance inequality*, we define the inner product

$$\langle u, v \rangle \equiv \mathbb{E}[uv] = \int\int \mathrm{d}u\mathrm{d}v\ P(u, v)uv$$

(one can easily verify that this is indeed a proper inner product - it is symmetric, linear, and positive definite) such that $\mathbb{E}[XY] \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$.

variables $X$ and $Y$, $\langle XY\rangle^2 \leq \langle X^2\rangle\langle Y^2\rangle$, so

$$(b'(s_0)+1)^2 \leq \left\langle (\hat{s}(x) - \langle \hat{s}(x)\rangle)^2 \right\rangle \left\langle \left(\frac{\mathrm{d}}{\mathrm{d}s_0}\log P(x|s_0)\right)^2 \right\rangle = \mathrm{Var}[\hat{s}(x)]J(s_0)$$

$$\Leftrightarrow \mathrm{Var}[\hat{s}(x)] \geq \frac{(b'(s_0)+1)^2}{J(s_0)}$$

Thus, the Fisher information $J(s_0)$ contributes to a lower bound on the variance of any estimator. Particularly, if the estimator is unbiased (in which case $b'(s_0) = 0$), the inverse Fisher information exactly gives that lower bound.

## 5.3 Zhang & Sejnowski (1999): Optimal Tuning Curve Widths

Consider a population of neurons $a = 1, \ldots, N$ coding for a stimulus $\mathbf{s} \in \mathbb{R}^D$, with homogenous tuning curves given by

$$f_a(\mathbf{s}) = r_{max}\phi\left(\frac{\sum_{d=1}^{D}(s_d - c_d^a)^2}{\sigma^2}\right) = r_{max}\phi\left(\frac{\xi^a}{\sigma^2}\right)$$

where $c_d^a$ are the tuning curve center in each dimension, and $\phi(\cdot)$ is a monotonically decreasing function (e.g. $\phi(x) = e^{-x}$ for Gaussian tuning curve). Note that because we enforce the tuning curve widths $\sigma$ to be the same for each stimulus dimension, the resulting tuning curves $f_a(\mathbf{s})$ are circularly symmetric. We then assume that each neuron's response is independently distributed according to a probability distribution dependent on its tuning, i.e. $P(\mathbf{r}|\mathbf{s}) = \prod_a P(r_a|f_a(\mathbf{s}))$. The Fisher information for the population is then given by the sum of the Fisher information for each neuron:

$$\begin{aligned}
\mathbf{J}(\mathbf{s}) &= \left\langle -\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\log P(\mathbf{r}|\mathbf{s}) \right\rangle \\
&= \left\langle -\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\sum_a \log P(r_a|f_a(\mathbf{s})) \right\rangle \\
&= \sum_a \left\langle -\frac{\mathrm{d}^2}{\mathrm{d}\mathbf{s}^2}\log P(r_a|f_a(\mathbf{s})) \right\rangle \\
&= \sum_a \mathbf{J}^{(a)}(\mathbf{s})
\end{aligned}$$

The question we ask now is: what tuning curve width $\sigma$ maximizes the population Fisher information?

We begin by deriving the Fisher information matrix for a single neuron $a$. Using the squared first derivative form, we have

$$J_{ij}^{(a)}(\mathbf{s}) = \left\langle \frac{\mathrm{d}}{\mathrm{d}s_i}\log P(r_a|f_a(\mathbf{s}))\frac{\mathrm{d}}{\mathrm{d}s_j}\log P(r_a|f(\mathbf{s})) \right\rangle$$

Applying chain rule, we can get an expression for the two derivatives:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s_i}\log P(r_a|f_a(\mathbf{s})) &= \frac{1}{P(r_a|\mathbf{s})}\frac{\partial P(r_a|f_a(\mathbf{s}))}{\partial s_i} \\
&= \frac{1}{P(r_a|\mathbf{s})}\frac{\partial P(r_a|f_a(\mathbf{s}))}{\partial f_a(\mathbf{s})}\frac{\partial f_a(\mathbf{s})}{\partial \phi(\xi^a/\sigma^2)}\frac{\partial \phi(\xi^a/\sigma^2)}{\partial \xi^a}\frac{\partial \xi^a}{\partial s_i} \\
&= \frac{1}{P(r_a|\mathbf{s})}\frac{\partial P(r_a|f_a(\mathbf{s}))}{\partial f_a(\mathbf{s})} r_{max}\frac{\phi'(\xi^a/\sigma^2)}{\sigma^2} 2(s_i - c_i^a)
\end{aligned}$$

We thus have:

$$J_{ij}^{(a)}(\mathbf{s}) = K_a(\xi_a)\frac{(s_i - c_i^a)(s_j - c_j^a)}{\sigma^4}$$

where

$$K_a(\xi_a) = \left\langle 4\left(\frac{1}{P(r_a|\mathbf{s})}\frac{\partial P(r_a|f_a(\mathbf{s}))}{\partial f_a(\mathbf{s})} r_{max}\phi'(\xi^a/\sigma^2)\right)^2 \right\rangle$$

is the only term where the expectation appears since it is the only term dependent on the activity $r_a$ over which the expectation is defined. Defining

$$\xi_i^a = \frac{s_i - c_i}{\sigma}$$

such that $\xi_a = \sigma^2 \sum_i {\xi_i^a}^2$, we note that since $\phi(\xi^a/\sigma^2)$ is a monotonically decreasing function of ${\xi_i^a}^2$, it is symmetric around $\xi_i^a = 0$. Thus, $\phi'(\xi^a/\sigma^2)^2$ and $f_a(\mathbf{s})$ are as well, so $K_a(\xi_a)$ is also symmetric around $\xi_i^a = 0$.

We now assume that the neuron tuning centers are uniformly distributed across the stimulus space, with $P(c_i^a) = p_c$. Taking the limit of $N \to \infty$ to approximate sums with integrals, we can then write

$$\begin{aligned} J_{ij}(\mathbf{s}) &= \sum_a J_{ij}^{(a)}(\mathbf{s}) \\ &\approx \int \mathrm{d}c_1^a \int \mathrm{d}c_2^a \ldots \int \mathrm{d}c_D^a \; p_c J_{ij}^{(a)}(\mathbf{s}) \\ &\approx \int \mathrm{d}c_1^a \int \mathrm{d}c_2^a \ldots \int \mathrm{d}c_D^a \; p_c K_a(\xi_a)^2 \frac{(s_i - c_i^a)(s_j - c_j^a)}{\sigma^4} \end{aligned}$$

To evaluate this integral, we exploit the fact that $K_a(\xi_a)$ is symmetric around $\xi_i^a = 0$ by performing the change of variables $c_i^a \to \xi_i^a$, such that $\mathrm{d}c_i^a = -\sigma \mathrm{d}\xi_i^a$:

$$\begin{aligned} J_{ij}(\mathbf{s}) &\approx \frac{1}{\sigma^2} \int \sigma \mathrm{d}\xi_1^a \int \sigma \mathrm{d}\xi_2^a \ldots \int \sigma \mathrm{d}\xi_D^a \; p_c K_a(\xi_a) \xi_i^a \xi_j^a \\ &= \frac{\sigma^D}{\sigma^2} \int \mathrm{d}\xi_1^a \int \mathrm{d}\xi_2^a \ldots \int \mathrm{d}\xi_D^a \; p_c K_a(\xi^a) \xi_i^a \xi_j^a \end{aligned}$$

Because $K_a(\xi^a)$ is symmetric around $\xi_i^a = 0$, we have that

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{d}\xi_i^a \; p_c K_a(\xi^a) \xi_i^a &= \int_{-\infty}^{0} \mathrm{d}\xi_i^a \; p_c K_a(\xi^a) \xi_i^a + \int_0^{\infty} \mathrm{d}\xi_i^a \; p_c K_a(\xi^a) \xi_i^a \\ &= -\int_0^{\infty} \mathrm{d}\xi_i^a \; p_c K_a(-\xi^a) \xi_i^a + \int_0^{\infty} \mathrm{d}\xi_i^a \; p_c K_a(\xi^a) \xi_i^a \\ &= -\int_0^{\infty} \mathrm{d}\xi_i^a \; p_c K_a(\xi^a) \xi_i^a + \int_0^{\infty} \mathrm{d}\xi_i^a \; p_c K_a(\xi^a) \xi_i^a = 0 \end{aligned}$$

Therefore, when $i \neq j$

$$J_{ij}(\mathbf{s}) \approx \frac{\sigma^D}{\sigma^2} \int \mathrm{d}\xi_1^a \ldots \int \mathrm{d}\xi_{i-1}^a \int \mathrm{d}\xi_{i+1}^a \ldots \int \mathrm{d}\xi_D^a \; \xi_j^a \int \mathrm{d}\xi_i^a \; p_c K_a(\xi^a) \xi_i^a = 0$$

When $i = j$, on the other hand, we have

$$J_{ii}(\mathbf{s}) \approx \frac{\sigma^D}{\sigma^2} \int \mathrm{d}\xi_1^a \int \mathrm{d}\xi_2^a \ldots \int \mathrm{d}\xi_D^a \; p_c K_a(\xi^a) {\xi_i^a}^2 = \sigma^{D-2} A$$

where $A$ is independent of $\sigma$.

We have thus found that the total Fisher information in the population is proportional to $\sigma^{D-2}$, where $D$ is the dimensionality of the stimulus. This yields a few surprising results regarding the optimal tuning curve width $\sigma$:

- If $D = 1$, you want the tuning curves as narrow as possible, down to the smallest resolution between neighboring $c_i^a$ (but not smaller than this).
- If $D = 2$, the Fisher information is completely independent of the tuning curve widths.
- If $D > 2$, the wider the tuning width the better. Optimality is achieved when the width of the tuning curve spans the stimulus space.

Furthermore, it turns out that if you allow the tuning curve widths to vary between stimulus dimensions (i.e. allow $f_a(\mathbf{s})$ to not be circularly symmetric), then maximizing the Fisher information gives you a cartesian code whereby the optimal tuning curve width is narrow in some dimensions and wide in others.

### 5.4 Olshausen & Field (1996): Sparse Coding

### 5.5 Correlations and Population Coding

### 5.6 Coding Uncertainty

DDPCs

## 6 Suggested Papers

- Network dynamics
  - [Wilson and Cowan, 1972]
  - [Hopfield, 1982]
  - [Seung, 1996]
  - [Latham et al., 2000]
  - Mean-field analysis:
    - [Sompolinsky et al., 1988]
    - [Vreeswijk and Sompolinsky, 1998]
    - [Renart et al., 2010]
    - [Rosenbaum et al., 2017]
    - [Mastrogiuseppe and Ostojic, 2017]
- Correlations and information
  - [Zohary et al., 1994]
  - [Abbott and Dayan, 1999]
  - [Moreno-Bote et al., 2014]
  - [Kohn et al., 2016]
- Coding
  - [Olshausen and Field, 1996]
  - [Zhang and Sejnowski, 1999]
- Optimality
  - [Ernst and Banks, 2002]
  - [Kording and Wolpert, 2004]
- Characterizing neural responses
  - [Simoncelli et al., 2004]

# 7 Appendices

## 7.1 Important Constants In Neuroscience

| Symbol | Value | Name | Notes |
|---|---|---|---|
| - | $10^{11}$ | number of neurons | in human brain |
| $K$ | 1000 | avg connections per neuron | in cortex? (PEL) |
| $V_{rest}$ | $-70$mV | resting membrane potential | |
| $V_{th}$ | $-50$mV | spiking threshold | in reality not fixed |
| $E_{Na}$ | 50mV | $Na^+$ reversal potential | |
| $E_K$ | $-90-70$mV | $K^+$ reversal potential | |
| $E_{Cl}$ | $-65-60$mV | $Cl^-$ reversal potential | |
| $E_{Ca}$ | 150mV | $Ca^2+$ reversal potential | |
| $\tau_m$ | $10-100$ms | membrane time constant | $\tau_m = c_m r_m = C_m R_m$, independent of membrane surface area |
| $r_m$ | 1MΩmm$^2$ | specific membrane resistance | membrane resistance of a neuron with surface area $A$ given by $R_m = \frac{r_m}{A}$, $r_m$ varies with $V$ |
| $c_m$ | 10nF/mm$^2$ | specific membrane capacitance | membrane capacitance of a neuron with surface area $A$ given by $C_m = c_m A$ |
| $A$ | .01mm$^2$ | neuronal surface area | |
| $r_L$ | 1kΩmm | intracellular resistivity | a property of cell cytoplasm; longitudinal resistance in a neurite of length $L$ and cross-sectional radius $a$ is given by $R_L = \frac{L \times r_L}{\pi a^2}$ |
| a | $2\mu$m | cross-sectional radius of a dendrite | velocity of signal propagation in axon, dendrite scales with $a$, $\sqrt{a}$ |
| $\lambda$ | 1mm | electrotonic length of a dendrite | $\lambda = \sqrt{\frac{r_m a}{2 r_L}}$, sets the scale of spatial decay of a constant current injection (for infinite length dendrite) $\rightarrow$ dendrites can't be much longer than this |
| $R_\lambda$ | ∼MΩ | input resistance | $R_\lambda = \frac{r_m}{\lambda 2\pi a}$, the ratio of equilibrium potential to injected current (for constant current injection in infinite cable) |
| $g_i^{open}$ | 25pS | open channel conductance | |
| - | 1mV | EPSP | |
| - | 1ms | PSP rise time constant | in CA3 pyramidal cell |
| - | 5ms | PSP decay time constant | in CA3 pyramidal cell |

## 7.2 Useful Approximations and Maths Facts

- $\log(1+z) \approx z$ for small $z$
- $(1+z/n)^n \rightarrow e^z$ for big $n$
- $\tanh(z) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

## 7.3 Electrical Circuits

| Name | Symbol | Units |
|---|---|---|
| Charge | Q | $C = 6.2 \times 10^{18}$ electrons (coulombs) |
| Current | I | $A = C/s$ (amps = coulombs per second) |
| Voltage | V | $V = J/C$ (volts = potential energy joules per coulomb) |
| Resistance | R | $\Omega = V/A$ (ohms = volts per amp) |
| Capacitance | C | $F = C/V$ (farad = coulomb per volt) |

We therefore have:

$$\begin{aligned} I &= \frac{\mathrm{d}Q}{\mathrm{d}t} \\ CV = Q \Leftrightarrow C\frac{\mathrm{d}V}{\mathrm{d}t} &= I \\ R = \frac{I}{V} \Leftrightarrow \Delta V &= IR \quad \text{(Ohm's Law)} \end{aligned}$$

We can interpret these latter two equations as telling us that

- capacitance $C$ determines how much current $I$ is needed to change the voltage *at a given rate* $\frac{\mathrm{d}V}{\mathrm{d}t}$
- resistance $R$ determines how much current $I$ is needed to change the volltage *by a given amount* $\Delta V$

## 7.4 Solving Differenetial Equations

### 7.4.1 First-Order ODEs: Method of Integrating Factors

Consider a differential equation of the form:

$$\frac{\mathrm{d}y}{\mathrm{d}x} + p(x)y(x) = g(x)$$

Because of the $y(x)$ term isolated from the derivative on the left-hand side, we can't solve this by straight-forward integration. We can deal with this, however, via the product rule by multiplying both sides by the *integrating factor*

$$v(x) = \int p(x)\mathrm{d}x \Leftrightarrow \frac{\mathrm{d}v}{\mathrm{d}x} = p(x)$$

which gives us:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}y(x)e^{v(x)} &= \frac{\mathrm{d}y}{\mathrm{d}x}e^{v(x)} + \frac{\mathrm{d}v}{\mathrm{d}x}y(x)e^{v(x)} \\ &= \left(\frac{\mathrm{d}y}{\mathrm{d}x} + p(x)y(x)\right)e^{v(x)} \\ &= g(x)e^{v(x)} \end{aligned}$$

such that we can now simply integrate both sides to get our solution:

$$\begin{aligned} \int \frac{\mathrm{d}}{\mathrm{d}x}y(x)e^{v(x)}\mathrm{d}x &= \int g(x)e^{v(x)}\mathrm{d}x \\ \Leftrightarrow y(x) &= e^{-v(x)}\int g(x)e^{v(x)}\mathrm{d}x \end{aligned}$$

### 7.4.2 Homogenous Second-Order ODEs

Consider a differential equation of the form

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + p\frac{\mathrm{d}y}{\mathrm{d}x} + qy(x) = 0$$

with $p, q$ constant coefficients. Let $y'' = \frac{\mathrm{d}^2y}{\mathrm{d}x^2}, y' = \frac{\mathrm{d}y}{\mathrm{d}x}$. We now note that if we can find a pair $a, b$ such that $p = -(a+b), q = ab$, we can turn this homoegenous second-order ODE into a homogenous first-order ODE:

$$\begin{aligned} y'' + py' + qy &= y'' - (a+b)y' + aby \\ &= (y' - ay)' + b(ay - y') \\ &= 0 \\ \Leftrightarrow (y' - ay)' &= b(y' - ay) \end{aligned}$$

Making the substitution $u = y' - ay$ and solving the resulting first-order ODE we then have:

$$\begin{aligned} u' &= bu \\ \Leftrightarrow u(x) &= Ce^{bx} \\ \Rightarrow y' - ay &= Ce^{bx} \\ \Leftrightarrow (ye^{-ax})' &= Ce^{(b-a)x} \\ \Leftrightarrow y(x) &= c_1 e^{ax} + c_2 e^{bx} \end{aligned}$$

Unless $a = b$, in which case the fourth line becomes

$$\begin{aligned} (ye^{-ax})' &= C \\ \Leftrightarrow y(x) &= e^{ax}(c_1 x + c_2) \end{aligned}$$

So all we need to do to solve a homogenous second-order ODE with constant coefficients $p, q$ is to find $a, b$ such that $p = -(a + b), q = ab$. We can do this easily by noting that $a, b$ are the solutions to the quadratic equation

$$r^2 + pr + q = r^2 - (a + b)r + ab = (r - a)(r - b) = 0$$

We call this equation the *characteristic equation* of the above second-order ODE. Using the quadratic formula, we then have:

$$a, b = \frac{-p \pm \sqrt{p^2 - 4q}}{2}$$

### 7.4.3 Nth-order Inhomogenous ODEs: Green's Function

Consider a differential equation of the form

$$\frac{\mathrm{d}^n y}{\mathrm{d}x^n} + \frac{\mathrm{d}^{n-1} y}{\mathrm{d}x^{n-1}} + \ldots + \frac{\mathrm{d}y}{\mathrm{d}x} + y(x) = g(x)$$

Recalling that differentiation is a linear operation, we can define the linear operator $L$:

$$Ly(x) = \frac{\mathrm{d}^n y}{\mathrm{d}x^n} + \ldots + \frac{\mathrm{d}y}{\mathrm{d}x} + y(x)$$

We now find a function $G(x, s)$ such that

$$LG(x, s) = \delta(s - x)$$

This is called a *Green's function*, which depends on the linear operator $L$.

Once we have found the Green's function, we can use it to solve the differential equation by noting that

$$\int LG(x, s)g(s)\mathrm{d}s = g(x)$$

Crucially, since $L$ is a linear operator with respect to $x$ (not $s$), we can pull it out of the integral. We can thus rewrite the differential equation as:

$$\begin{aligned} Ly(x) &= g(x) \\ &= \int LG(x, s)g(s)\mathrm{d}s \\ &= L \int G(x, s)g(s)\mathrm{d}s \\ \Leftrightarrow y(x) &= \int G(x, s)g(s)\mathrm{d}s \end{aligned}$$

which will hopefully be an easy integral if the Green's function $G(x, s)$ is of a nice form.

## 7.5 Dynamical Systems Analysis

Consider a dynamical system of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x,y)$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = g(x,y)$$

Since the dynamics depend only on the variables $x, y$ themselves and nothing else, we call such a system *autonomous*. To understand this system, we would like to know the behavior of trajectories of $(x(t), y(t))$ over time. We can gain such an understanding qualitatively by plotting the *nullclines* of the system in the $x-y$ plane, given by

$$f(x,y) = 0$$
$$g(x,y) = 0$$

We can then construct the so-called *phase plane* by sketching trajectories in the $x-y$ plane. For simple systems, we can directly calculate trajectories by picking an initial condition and solving the differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{f(x,y)}{g(x,y)}$$

However, this is usually impossible to do analytically, so we instead turn to the nullclines to guide us via the following rules:

- Trajectories can only cross the $x-$nullcline $f(x,y) = 0$ vertically (i.e. with $\frac{\mathrm{d}x}{\mathrm{d}t} = 0$)
- Trajectories can only cross the $y-$nullcline $g(x,y) = 0$ horizontally (i.e. with $\frac{\mathrm{d}y}{\mathrm{d}t} = 0$)
- Regions enclosed by the nullclines have $\frac{\mathrm{d}x}{\mathrm{d}y}$ with constant sign
- Crossings of the two nullclines are fixed points (stable/unstable) of the system

This last point is of great importance, as often what we are most interested in is the long-run behavior of the system. Thus, we would like to be able to know the behavior of the system near each of the fixed points. We can do so via standard stability analysis. Consider a fixed point $(x^*, y^*)$ of the above system given, found by solving the equation

$$f(x^*, y^*) = g(x^*, y^*) = 0$$

To understand the system's behavior near this point, we analyze the dynamics at a nearby point

$$(\tilde{x}(t), \tilde{y}(t)) = (x^* + \delta x(t), y^* + \delta y(t))$$

to examine where it ends up in the limit of $t \to \infty$. If $(\tilde{x}(t), \tilde{y}(t)) \to (x^*, y^*)$, i.e. $(\delta x(t), \delta y(t)) \to (0,0)$, as $t \to \infty$ then we know the fixed point $(x^*, y^*)$ is stable.

Assuming $(\delta x(t), \delta y(t))$ to be very small, we can safely approximate the dynamics at $(\tilde{x}, \tilde{y})$ to 1st order:

$$\frac{\mathrm{d}\tilde{x}}{\mathrm{d}t} = \frac{\mathrm{d}\delta x}{\mathrm{d}t} = f(x^* + \delta x, y^* + \delta y) \approx f(x^*, y^*) + f_x(x^*, y^*)\delta x + f_y(x^*, y^*)\delta y$$
$$\frac{\mathrm{d}\tilde{y}}{\mathrm{d}t} = \frac{\mathrm{d}\delta y}{\mathrm{d}t} = g(x^* + \delta x, y^* + \delta y) \approx g(x^*, y^*) + g_x(x^*, y^*)\delta x + g_y(x^*, y^*)\delta y$$

where I have used the notation $f_z(a,b) = \left.\frac{\partial f}{\partial z}\right|_{x=a, y=b}$. Since $f(x^*, y^*) = g(x^*, y^*) = 0$, we can rewrite this approximation in matrix notation as follows:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{J}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} \delta x \\ \delta y \end{bmatrix}$$

and

$$\mathbf{J} = \begin{bmatrix} f_x(x^*, y^*) & f_y(x^*, y^*) \\ g_x(x^*, y^*) & g_y(x^*, y^*) \end{bmatrix}$$

is the Jacobian of the vector-valued function $\mathbf{f}(x, y) = \begin{bmatrix} f(x, y) & g(x, y) \end{bmatrix}^{\mathrm{T}}$, evaluated at $(x^*, y^*)$.

We now have a linear dynamical system that we can actually solve. Note what we have done: by picking a point very near to the fixed point and approximating its dynamics to first-order, we have effectively *linearized* the dynamics around this fixed point, giving us a linear system that we can analyze and solve. Specifically, the solution to this linear system is given by

$$\mathbf{x}(t) = c_1 e^{\lambda_1 t}\mathbf{v}_1 + c_2 e^{\lambda_2 t}\mathbf{v}_2$$

where $\lambda_1, \lambda_2$ and $\mathbf{v}_1, \mathbf{v}_2$ are the eigenvalues and eigenvectors of the $2 \times 2$ Jacobian matrix $\mathbf{J}$. Thus, if $\mathrm{Re}(\lambda_1), \mathrm{Re}(\lambda_2) < 0$, we know that $e^{\lambda_1 t}, e^{\lambda_2 t} \to 0$ and therefore $(\delta x, \delta y) \to 0$ as $t \to \infty$, so we canc conclude $(x^*, y^*)$ is a stable fixed point. Otherwise, $(x^*, y^*)$ could be either unstable, a saddle node, or a limit cycle (see table 7.5). We therefore need only calculate the eigenvalues of the Jacobian matrix $\mathbf{J}$ to determine qualitative behavior around the fixed point $(x^*, y^*)$:

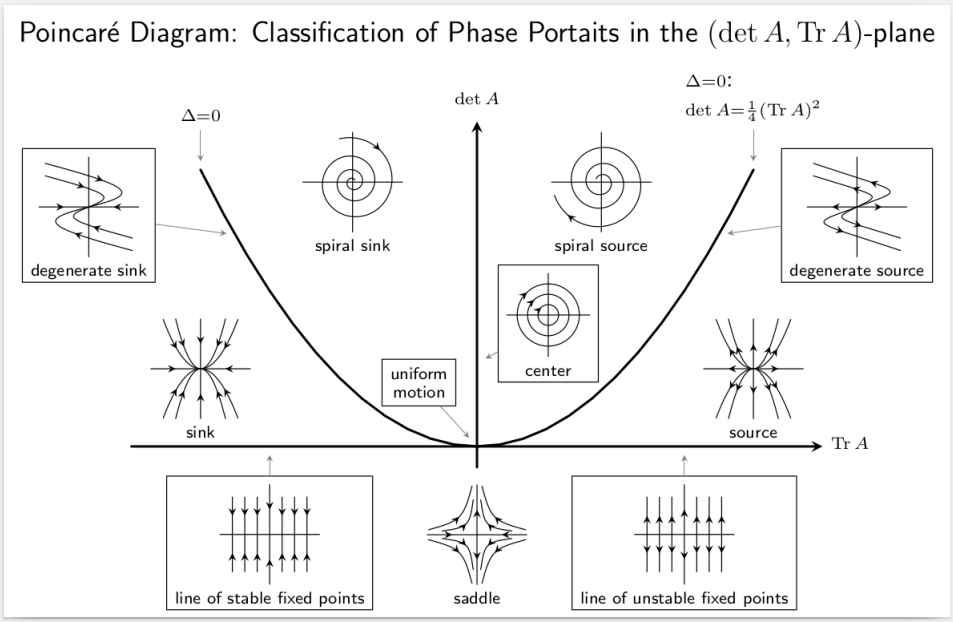
$$\begin{aligned} & \mathbf{J}\mathbf{v} = \lambda\mathbf{v} \\ \Leftrightarrow\ & (\mathbf{J} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0} \\ \Rightarrow\ & |\mathbf{J} - \lambda\mathbf{I}| = \mathbf{0} \quad \text{for non-zero } \mathbf{v} \\ \Leftrightarrow\ & \lambda^2 - \underbrace{\mathrm{Tr}[\mathbf{J}]}_{T} + \underbrace{|\mathbf{J}|}_{D} = 0 \\ \Rightarrow\ & \lambda_\pm = \frac{T \pm \sqrt{T^2 - 4D}}{2} \end{aligned}$$

where the third line follows from the fact that, for there to be a non-zero vector $\mathbf{v}$ that satisfies the equation in the second line, the matrix $\mathbf{J} - \lambda\mathbf{I}$ must have a non-zero nullspace and therefore not be full-rank, which implies that its determinant must be 0. We can thus easily derive the following conditions for stability of the fixed point:

$$\begin{aligned} T &< 0 \\ D &> 0 \end{aligned}$$

The full picture is given by table 7.5 and figure 7.5 below.

| Fixed point | $\mathrm{Tr}[\mathbf{J}]$ | $\mathrm{Det}[\mathbf{J}]$ | Real part | Imaginary part |
|---|---|---|---|---|
| stable node | $T < 0$ | $T^2 > 4D > 0$ | $\mathrm{Re}(\lambda_\pm) < 0$ | $\mathrm{Im}(\lambda_\pm) = 0$ |
| stable spiral | $T < 0$ | $4D > T^2 > 0$ | $\mathrm{Re}(\lambda_\pm) < 0$ | $\mathrm{Im}(\lambda_\pm) \neq 0$ |
| unstable node | $T > 0$ | $T^2 > 4D > 0$ | $\mathrm{Re}(\lambda_\pm) > 0$ | $\mathrm{Im}(\lambda_\pm) = 0$ |
| unstable spiral | $T > 0$ | $4D > T^2 > 0$ | $\mathrm{Re}(\lambda_\pm) > 0$ | $\mathrm{Im}(\lambda_\pm) \neq 0$ |
| center (limit cycle??) | $T = 0$ | $D > 0$ | $\mathrm{Re}(\lambda_\pm) = 0$ | $\mathrm{Im}(\lambda_\pm) \neq 0$ |
| saddle | - | $D < 0$ | $\mathrm{Re}(\lambda_+) > 0 > \mathrm{Re}(\lambda_-)$ | $\mathrm{Im}(\lambda_\pm) = 0$ |
| star/degenerate node | $T^2 = 4D$ | $D \geq 0$ | $\mathrm{Re}(\lambda_+) = \mathrm{Re}(\lambda_-)$ | $\mathrm{Im}(\lambda_\pm) = 0$ |




## 7.6 Fourier Transform

Given a function $f(x)$ in space or time (i.e. $x$ in cm or seconds), one can equialently express it in the frequency domain via its Fourier transform $F(\omega)$:

$$F(\omega) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i \omega x} \mathrm{d}x$$
$$f(x) = \int_{-\infty}^{\infty} F(\omega) e^{2\pi i \omega x} \mathrm{d}\omega$$

When the units of $\omega$ don't matter to the given derivation, the $2\pi$ can be dropped. One must then simply rescale $\omega \to \frac{\omega}{2\pi}$ to interpret it as a frequency in inverse units of $x$ (e.g. Hz for $x$ in seconds).

Some important Fourier transforms to know are given in the table below, where $\omega > 0$ is in inverse units of $x$:

| $f(x)$ | $F(\omega)$ |
|---|---|
| $\sin(kx)$ | $\delta(\omega - 2\pi k)$ |
| $\delta(x)$ | $1$ |
| $\frac{d^n}{dx^n} g(x)$ | $(2\pi i\omega)^n G(\omega)$ |
| $e^{-ax^2}, a > 0$ | $\sqrt{\frac{\pi}{a}} e^{-\frac{\pi^2\omega^2}{a}}$ |
| $\Theta(x) e^{-ax}, a > 0$ | $\frac{1}{2\pi i\omega + a}$ |

A useful property of the Fourier transform is the so-called *Convolution Theorem*, which states that the Fourier transform of the convolution of two functions is equal to the product of their Fourier transforms. Let

$$h(x) = \int f(x') g(x - x') \mathrm{d}x'$$

be the convolution of $f(\cdot)$ and $g(\cdot)$. Then its Fourier transform is:

$$\begin{aligned} H(\omega) &= \int h(x) e^{-2\pi i \omega x} \mathrm{d}x \\ &= \int \int f(x') g(x - x') \mathrm{d}x' e^{-2\pi i \omega x} \mathrm{d}x \\ &= \int f(x') \int g(x - x') e^{-2\pi i \omega x} \mathrm{d}x \mathrm{d}x' \\ &\overset{y=x-x'}{=} \int f(x') \int g(y) e^{-2\pi i \omega (y + x')} \mathrm{d}y \mathrm{d}x' \end{aligned}$$

$$= \int f(x')e^{-2\pi i\omega x'}\mathrm{d}x' \int g(y)e^{-2\pi i\omega y}\mathrm{d}y$$
$$= F(\omega)G(\omega)$$

where in the third line we switched the order of integration[15] and in the fourth line we made the substitution $y = x - x'$.

## 7.7 Central Limit Theorem

The Central Limit Theorem states that for any set of *independent* 0-mean random variables $X_1, X_2, \ldots, X_n$ with variances given by $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, in the limit of $n \to \infty$

$$P\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}\sigma} < C\right) = P(Z_n < C) \to P(Z < C), \quad Z \sim \mathcal{N}(0,1)$$

where

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^n \sigma_i^2$$

The arrow here means that the cumulative distribution of the random variable $Z_n$ *converges in distribution* to that of a standard Gaussian. This means that different segments of the distribution may converge to Gaussianity at different rates (e.g. the tails of its distribution will converge more slowly).

We prove it here via the *moment generating function* of a random variable $X$:

$$M_X(t) \equiv \mathbb{E}[e^{tX}]$$

Noting that this implies

$$M_X(t) = \mathbb{E}\left[1 + tX + \frac{1}{2!}t^2X^2 + \frac{1}{3!}t^3X^3 + ...\right] = 1 + t\mathbb{E}[X] + \frac{1}{2!}t^2\mathbb{E}[X^2] + \frac{1}{3!}t^3\mathbb{E}[X^3] + ...$$

it is easy to see that the following holds:

$$\frac{\mathrm{d}^\ell M_X}{\mathrm{d}t^\ell}\Big|_{t=0} = \mathbb{E}[X^\ell]$$

Thus its name.

We then require four facts:

1. The moment generating function of a sum of independent random variables $Z = X + Y$ is the product of their moment generating functions $M_X(t), M_Y(t)$:

$$M_Z(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = M_X(t)M_Y(t)$$

2. The moment generating function of a linear transformation of a random variable $Z = aX + b$ is given by:

$$M_Z(t) = \mathbb{E}[e^{atX+bt)}] = e^{bt}\mathbb{E}[e^{atX}] = e^{bt}M_X(at)$$

3. If the moment generating functions $M_{X_1}(t), M_{X_2}(t), \ldots$ of a sequence of random variables $X_1, X_2, \ldots$ converge to some moment generating function $M_X(t)$, i.e.

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$$

then their respective cumulative density functions $F_1(x), F_2(x), \ldots$ converge in distribution to the cumulative density function $F(x)$ of $X$:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

[15] This is allowed under certain relatively soft constraints on $f(x), g(x)$, namely that the integral of their absolute value be finite, I believe (cf. Fubini's Theorem).

4. The moment generating function of a standard Gaussian random variable $X \sim \mathcal{N}(0,1)$ is given by:

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx - \frac{x^2}{2}} \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x^2 - 2tx)}{2}} \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2 - t^2}{2}} \mathrm{d}x \\ &= \frac{e^{\frac{t^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}} \mathrm{d}x \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

By points 3 and 4, then, to prove the CLT it suffices to show that

$$\lim_{n \to \infty} M_{Z_n}(t) = e^{\frac{t^2}{2}}$$

Using points 1 and 2, we have:

$$M_{Z_n}(t) = \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{\sqrt{n}\sigma}\right)$$

Expanding the individual moment generating functions, we have:

$$M_{X_i}\left(\frac{t}{\sqrt{n}\sigma}\right) = 1 + \frac{t}{\sqrt{n}\sigma}\mathbb{E}[X_i] + \frac{t^2}{2n\sigma^2}\mathbb{E}[X_i^2] + \frac{t^3}{3!n^{3/2}\sigma^3}\mathbb{E}[X_i^3] + \dots$$

As $n \to \infty$, the latter terms will go to 0 faster than the earlier terms, eventually giving us

$$M_{X_i}\left(\frac{t}{\sqrt{n}\sigma}\right) \to 1 + \frac{t^2\sigma_i^2}{2n\sigma^2}$$

where I have also substituted in $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \sigma_i^2$ (true for all $i$). Assuming the individual variances $\sigma_i^2$ are not too different from each other, $\sigma_i^2/\sigma^2 \sim \mathcal{O}(1)$ and we can ignore it. Thus,

$$M_{Z_n}(t) \to \left(1 + \frac{t^2}{2n}\right)^n \to e^{\frac{t^2}{2}}$$

There is a more rigorous way of proving this without any handwaving, but one can see that this definitely holds for the case where the $X_i$ have equal variance (i.e. when they are i.i.d.).

# References


[Abbott and Dayan, 1999] Abbott, L. F. and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101.

[Dayan and Abbott, 2001] Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience*, volume 806. Cambridge, MA: MIT Press.

[Ernst and Banks, 2002] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

[Gerstner et al., 2014] Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

[Grabska-Barwińska and Latham, 2014] Grabska-Barwińska, A. and Latham, P. E. (2014). How well do mean field theories of spiking quadratic-integrate-and-fire networks work in realistic parameter regimes? *Journal of computational neuroscience*, 36(3):469–481.

[Hopfield, 1982] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

[Kohn et al., 2016] Kohn, A., Coen-Cagli, R., Kanitscheider, I., and Pouget, A. (2016). Correlations and neuronal population information. *Annual review of neuroscience*, 39.

[Kording and Wolpert, 2004] Kording, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244.

[Latham, 2017] Latham, P. E. (2017). Correlations demystified. *Nature neuroscience*, 20(1):6.

[Latham et al., 2000] Latham, P. E., Richmond, B., Nelson, P., and Nirenberg, S. (2000). Intrinsic dynamics in neuronal networks. i. theory. *Journal of neurophysiology*, 83(2):808–827.

[Mastrogiuseppe and Ostojic, 2017] Mastrogiuseppe, F. and Ostojic, S. (2017). Intrinsically-generated fluctuating activity in excitatory-inhibitory networks. *PLoS computational biology*, 13(4):e1005498.

[Moreno-Bote et al., 2014] Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417.

[Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607.

[Renart et al., 2010] Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., and Harris, K. D. (2010). The asynchronous state in cortical circuits. *science*, 327(5965):587–590.

[Rosenbaum et al., 2017] Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nature neuroscience*, 20(1):107.

[Seung, 1996] Seung, H. S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344.

[Simoncelli et al., 2004] Simoncelli, E. P., Paninski, L., Pillow, J., Schwartz, O., et al. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, 3(327-338):1.

[Sompolinsky et al., 1988] Sompolinsky, H., Crisanti, A., and Sommers, H.-J. (1988). Chaos in random neural networks. *Physical review letters*, 61(3):259.

[Vreeswijk and Sompolinsky, 1998] Vreeswijk, C. v. and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. *Neural computation*, 10(6):1321–1371.

[Wilson and Cowan, 1972] Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24.

[Zhang and Sejnowski, 1999] Zhang, K. and Sejnowski, T. J. (1999). Neuronal tuning: To sharpen or broaden? *Neural Computation*, 11(1):75–84.

[Zohary et al., 1994] Zohary, E., Shadlen, M. N., and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140.