Gatsby Computational Neuroscience Unit
Theoretical Neuroscience

Final examination, theoretical neuroscience
5 May 2023

Part II – long questions

There are four questions, one from each main section of the course. Please answer three out of the four, starting the answers for each question on a new page. Don't forget to write your name at the top of the answer to each question.

Good luck!

1 Biophysics

Suppose evolution hadn't invented an h-current. In that case, the equations describing a single neuron might look like

$$\tau \frac{dV}{dt} = -(V - \mathcal{E}_L) - \rho_m m_\infty(V)(V - \mathcal{E}_{Na}) - \rho_n n(V - \mathcal{E}_L)$$
$$\tau_n \frac{dn}{dt} = n_\infty(V) - n\,.$$

Both $m_\infty(V)$ and $n_\infty(V)$ are increasing functions of V, and $\mathcal{E}_L$, the leak reversal potential is less than $\mathcal{E}_{Na}$, the sodium reversal potential. As usual, $0 < n_\infty(v) < 1$. Normally the second $\mathcal{E}_L$ would be the potassium reversal potential (which is slightly less than $\mathcal{E}_L$); I'm using $\mathcal{E}_L$ to reduce the number of parameters.

Assume that

$$\rho_m = 5$$
$$\rho_n = 10\,.$$

1. The first thing we want to do is to get rid of as one of the parameters. Show that if we let $V = v + \mathcal{E}_L$, the equations become

$$\tau \frac{dv}{dt} = -v(1 + \rho_n n) + \rho_m m_\infty(v)(\Delta - v) \tag{3a}$$
$$\tau_n \frac{dn}{dt} = n_\infty(v) - n\,. \tag{3b}$$

   where

$$\Delta \equiv \mathcal{E}_{Na} - \mathcal{E}_L$$

   and we have abused notation somewhat and written $m_\infty(v)$ and $n_\infty(v)$ rather than the more accurate $m_\infty(v + \mathcal{E}_L)$ and $n_\infty(v + \mathcal{E}_L)$. We'll use $\Delta = 75$ mV (a fact you'll need below).

   (3 marks)

   Solution

   Completely straightforward algebra!

2. Define

$$I(v) \equiv -v(1 + \rho_n n) + \rho_m m_\infty(v)(\Delta - v)$$

   Plot $I(v)$ for several values of n, including $n = 0$. You need to use the fact that the $m_\infty(V)$ starts to activate around $V = -50$ mV (corresponding to $v = 15$), and plot the curve qualitatively. Alternatively, you can use the form of $m_\infty(v)$ given in problem 3, and plot $I(v)$ slightly more quantitatively.

   (7 marks)

Solution

For v near zero, $m_\infty(v)$ is negligible, and $I(v) = -v(1 + \rho_n n)$, which is a simple linear plot. As $m_\infty(v)$ starts to activate, around $v = 15$, $I(v)$ is pulled up by the $\Delta - v$ term. How high it's pulled up depends on n: when n is small it's pulled up enough to make $I(v)$ positive; when n is large, $I(v)$ remains negative as v increases. Typical plots, for a generic $m_\infty(v)$, are shown in Figure 1.

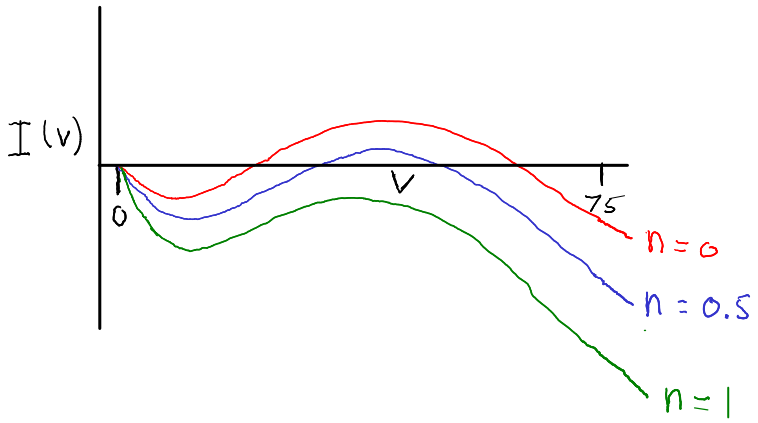

Figure 1: Typical plot of $I(v)$ for three different values of n.

3. Sketch the v-nullcline. For definiteness, let

$$m_\infty(v) = \frac{\mathrm{ReLU}(v - v_0)}{\alpha} = \frac{\max(0, v - v_0)}{\alpha} \tag{4}$$

where

$$v_0 = 15 \text{ mV}$$
$$\alpha = 15\,.$$

This isn't especially realistic, but it has the right flavor: it starts to activate at $v = 15$, which, assuming a resting membrane potential of -65 mV, corresponds to -50 mV; that's similar to the real m-current. What's not realistic is that m_∞ exceeds 1. However, the cost of a more complicated m_∞ is more algebra, so we'll stick to the simple version.

(10 marks)

Solution

To construct the nullclines, we plot $I(v)$ for several values of n and look for zeros. Using Eq. (4) for $m_\infty(v)$, we see that $I(v) = -v(1 + \rho_n n)$ when $v < v_0$, which is a simple linear plot. For $v > v_0$, things are only slightly more complicated,

$$I(v) = -v(1 + \rho_n n) + (\rho_m/\alpha)(v - v_0)(\Delta - v)\,.$$

One approach is to solve for n algebraically. However, I find it much easier to locate the zeros qualitatively, using the shape of $I(v)$. When $v > v_0$, $I(v)$ has a quadratic maximum. Our main task is to determine when that maximum is above zero (because that's the regime

where there are three points on the v-nullcline for fixed n). That's easiest to do if we write $I(v)$ in the form

$$I(v) = \frac{\rho_m}{\alpha}\left[\left(\frac{v_0+\Delta-\alpha(1+\rho_n n)/\rho_m}{2}\right)^2 - v_0\Delta - \left(v - \frac{v_0+\Delta-\alpha(1+\rho_n n)/\rho_m}{2}\right)^2\right],$$

which you can verify with a few lines of algebra. Using $v_0 = 15$, $\alpha = 15$, $\Delta = 75$, $\rho_m = 5$ and $\rho_n = 10$, we have

$$v_0 + \Delta - \alpha(1+\rho_n n)/\rho_m = 90 - (3+30n) = 87 - 30n\,,$$

and so,

$$I(v) = \frac{\rho_m}{\alpha}\left[(43.5-15n)^2 - 1125 - (v-(43.5-15n))^2\right].$$

For the right hand side to have real roots (with $n > 0$), we need $15n < 43.5 - \sqrt{1125}$. Given that $\sqrt{1125}$ is about 33, this means there are real roots for n smaller than about 2/3. Thus, $v = 0$ is always a point on the nullcline, and there are three points when n is smaller than about 2/3 and only one point at $v = 0$ when n is larger than about 2/3. Typical plots of $I(v)$ are shown in Figure 2a and the resulting nullcline is shown in Figure 2b.

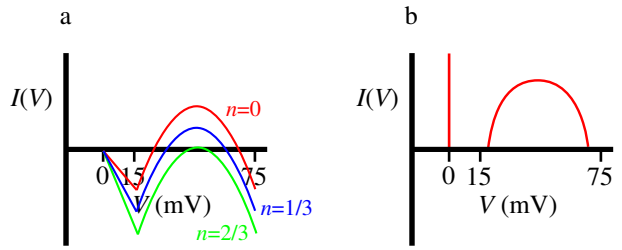

Figure 2: a) Plots of $I(v)$ for ReLU gain function. b) The resulting nullclines.

4. Sketch $n_\infty(v)$ using a shape that yields a single fixed point at $v = 0$. Assume that at $t = 0$, n is zero and v is set to a value just to the right of the middle zero of $I(v)$ (that is, $v(t=0) = v_1 + \epsilon$ where ϵ is small and v_1 is defined by $I(v_1) = 0$ and $dI(v_1)/dv_1 > 0$).

   Sketch the trajectories in v-n space in two limits: $\tau_n \gg \tau$ and $\tau_n \ll \tau$.

   (10 marks)

   Solution

   The blue curve shows the n-nullcline (which is given by $n_\infty(v)$) in a regime where the only fixed point is at $v = 0$.

   If $\tau_n \ll \tau$, movement in the n-direction is very fast relative to movement in the v-direction. In this case, the trajectory moves rapidly to the n-nullcline, which it crosses horizontally. It then moves (relatively) slowly along the n-nullcline until it reaches the equilibrium at $v = 0$. That's the green curve in Figure 3.

   If $\tau_n \gg \tau$, movement in the n-direction is very slow relative to movement in the v-direction. Thus, v quickly increases until it reaches the right branch of the v-nullcline. It then crosses that nullcline vertically, moves along it very slowly until it reaches the peak, and then moves more quickly back to $v = 0$. The resulting trajectory is shown in Fig. 3. That's the purple curve in Figure 3.

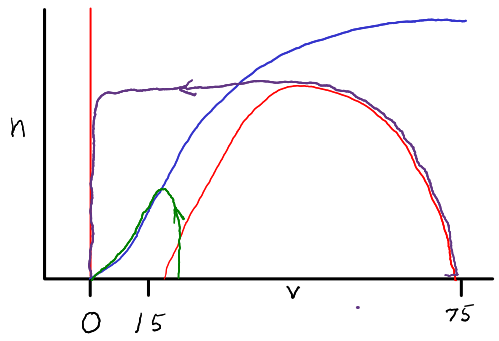

Figure 3: Nullclines (red and blue for v and n, respectively) in a regime where the only fixed point is at $n = v = 0$. Green trajectory τ_n is small, so n quickly relaxes to its nullcline, with little change in v. Small τ_n then causes the trajectory to stay near the n-nullcline. However, there is (relatively) slow drift in the v direction. Purple trajectory Things are pretty much reversed: rapid relaxation to the v-nullcline, and then relatively slow drift in the n-direction. However, we have a new phenomenon: there's a "catastrophe" at the peak of the v-nullcline, where the number of equilibria at fixed n go from three to one, causing rapid relaxation to the other other branch of the v-nullcline.

5. Explain why neither trajectory is ideal, from the point of view of biology. If you were designing a neuron like this to work in an actual brain, what considerations would go into choosing τ_n?

   (10 marks)

   Solution When the n-dynamics is fast ($\tau_n \ll \tau$; green trajectory), the membrane potential never gets very large, and spikes are small, which is not good – the membrane potential won't get high enough to activate the axons, so no signal will propagate. When the n-dynamics is slow ($\tau_n \gg \tau$; purple trajectory), spikes are large, but they're very slow. That not only slows down reaction time, it's also energetically expensive. The ideal case is somewhere in-between: τ_n should be small enough to avoid very slow spikes but large enough to give a reasonably large amplitude action potential.

2 Networks

Consider a model of binocular rivalry,

$$\tau \frac{du}{dt} = -u + \phi(I - wv) \tag{1a}$$

$$\tau \frac{dv}{dt} = -v + \phi(I - wu) \tag{1b}$$

where $w > 0$ and $\phi(\cdot)$ is positive, bounded, continuous, and a monotonic increasing function of its argument.

1. Show that there is exactly one symmetric fixed point (a solution with $u = v$ and $du/dt = dv/dt = 0$) always exists.

   (5 marks)

   Solution

   Define

   $$f(u) \equiv u - \phi(I - wu)\,.$$

   A symmetric fixed point occurs where $f(u) = 0$. Because ϕ is positive, we have $f(0) < 0$; because ϕ is bounded, we have $f(\infty) > 0$; and because ϕ is an increasing function of its argument, we have $df(u)/du > 0$. Thus, because ϕ is continuous, there must be one (and only 1) value of u at which $f(u) = 0$.

2. Let u^* be the symmetric fixed point. Show that it's stable if and only if $w\phi'(I - wu^*) < 1$.

   (10 marks)

   Solution

   Linearizing around $u = v = u^*$ (i.e., letting $u = u^* + \delta u$ and $v = v^* + \delta v$) gives us the eigenvalue equation

   $$\frac{d}{dt}\begin{pmatrix} \delta u \\ \delta v \end{pmatrix} = \begin{pmatrix} -1 & -w\phi'(I - wu^*) \\ -w\phi'(I - wu^*) & -1 \end{pmatrix}\begin{pmatrix} \delta u \\ \delta v \end{pmatrix}. \tag{2}$$

   The eigenvalues are both negative, meaning the fixed point is stable, if and only if two conditions are satisfied: the trace is negative (clearly satisfied) and the determinant is positive. The latter, denoted D, is given by

   $$D = 1 - w^2\phi'(I - wu^*)^2\,.$$

   For D to be positive, we must have $w^2\phi'(I - wu^*)^2 < 1$. Since both w and ϕ' are positive, this condition is equivalent to $w\phi'(I - wu^*) < 1$.

3. Assume that for some input current, say I^*, the symmetric fixed point, denoted u^*, is unstable. Sketch the generic shape of the largest eigenvalue of the linearized dynamics versus I. For this question, use a sigmoidal gain function: $\phi(z) = 1/(1 + \exp(-z))$. You want to consider a range of currents, from $I \ll I^*$ to $I \gg I^*$.

(15 marks)

Solution

Given Eq. (2), the largest eigenvalue, denoted λ, is given by

$$\lambda(I) = \frac{-2 + \sqrt{4 - 4\big(1 - w^2\phi'(I - wu^*(I))^2\big)}}{2} = w\phi'(I - wu^*(I)) - 1\,.$$

So we just need to plot $w\phi'(I - wu^*(I))$ versus I. Because ϕ is bounded from below and above and it's monotonic increasing, ϕ' must vanish for sufficiently small and large I. And we know that for $I = I^*$, $w\phi'(I - wu^*(I)) > 1$, indicating a positive eigenvalue. So $\lambda(I)$ should start negative for small (by which we really mean large and negative I), become positive for I near I^*, and be negative again for I large. The only question is: could it have multiple peaks? Because ϕ is sigmoidal,

$$\phi'(z) = \phi(z)(1 - \phi(z))\,,$$

which is a single-peaked function. So to determine whether or not there are multiple peaks, we just need to know how $I - wu^*(I)$ behaves versus I. To do that, we start with the equation for the equilibrium,

$$u^*(I) = \phi(I - wu^*(I))\,. \tag{3}$$

Differentiating with respect to I and rearranging terms gives

$$\frac{du^*(I)}{dI} = \frac{\phi'(I - u^*(I))}{1 + w\phi'(I - u^*(I))}\,.$$

Since the right hand side is non-negative, $u^*(I)$ must be an increasing function of I. Consequently, because of Eq. (3), $\phi(I - wu^*(I))$ must be an increasing function of I. And that in turn implies that $I - wu^*(I)$ is an increasing function of I. Thus, $\phi'(I - wu^*(I))$ has a single peak.

Putting that all together, a plot of $\lambda(I)$ versus I has the shape shown in Figure 4.

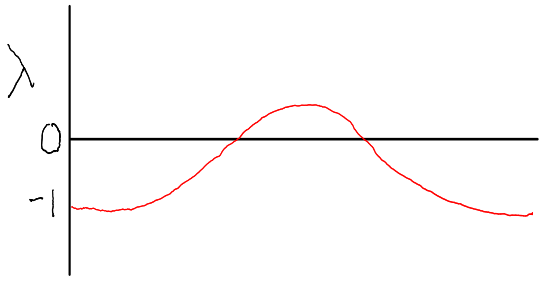

Figure 4: The largest eigenvalue of the dynamics linearized around a symmetric solution as a function of input current, I.

4. Sketch the nullclines in three regimes: a) $I \ll I^*$, b) $I = I^*$, and c) $I \gg I^*$. Again use a sigmoidal gain function, $\phi(z) = 1/(1 + \exp(-z)$.

(10 marks)

Solution

The extreme cases, $I \ll I^*$ and $I \gg I^*$, are the most straightforward: in the first case $\phi \approx 0$ and in the second $\phi \approx 1$. The nullclines are, therefore, nearly straight lines, with u and v either near zero or near 1. These are shown in Figures 5a and b, respectively.

The more interesting case is $I = I^*$. Although this is pretty straightforward as well: the u-nullcline is given by $u = \phi(I - wv)$ and the v-nullcline by $v = \phi(I - wu)$. These are shown in Figure 5c.

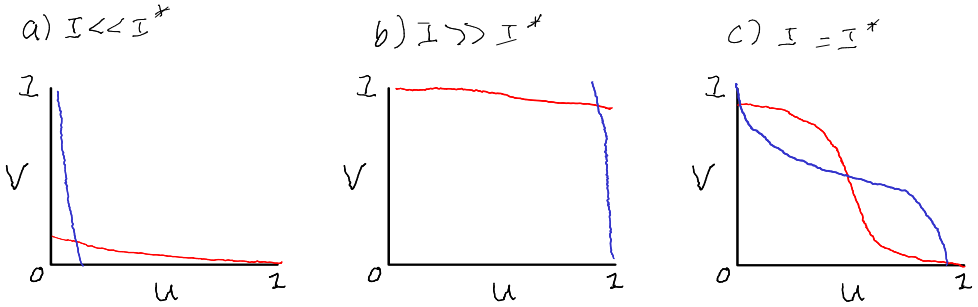

Figure 5: Nullclines for: a) $I \ll I^*$, b) $I \gg I^*$, and c) $I = I^*$. Red: v-nullclines. Blue: u-nullclines.

3 Coding & Population Dynamics

Consider a scenario in which a stimulus parametrised by scalar s is presented to an animal, resulting in a time-series of noisy feedforward inputs to cortex which we model as $\mathbf{u}(s,t) = \mathbf{g}(s) + \boldsymbol{\eta}(t)$. Here, $\mathbf{g}(s)$ is a time-independent but stimulus-dependent mean, and $\boldsymbol{\eta}(t)$ is an additive, stimulus-independent, multivariate normal noise process with zero mean and covariance $\Sigma_{\boldsymbol{\eta}}$.

First, consider an ideal observer that seeks to distinguish between two input signals (s_1, s_2), selecting the one that is more likely to have been presented given the sensory input.

1. Show that the optimal observer based on input activity at a single time t_0, will act by finding a linear projection $\mathbf{w}\cdot\mathbf{u}(s,t_0)$ for some vector $\mathbf{w}$, and comparing this value to a fixed threshold c. What is $\mathbf{w}$ in terms of the input properties?

   Solution We have

   $$\log p(s{=}s_i|\mathbf{u}(s,t_0)) = c_i - (\mathbf{u}(s,t_0) - \mathbf{g}(s_i))^{\mathsf{T}}\Sigma_{\boldsymbol{\eta}}^{-1}(\mathbf{u}(s,t_0) - \mathbf{g}(s_i))$$

   for a constant c_i. So

   $$\log \frac{p(s{=}s_2|\mathbf{u}(s,t_0))}{p(s{=}s_1|\mathbf{u}(s,t_0))} = c_2 - c_1 - \mathbf{g}(s_2)\Sigma_{\boldsymbol{\eta}}^{-1}\mathbf{g}(s_2) + \mathbf{g}(s_1)\Sigma_{\boldsymbol{\eta}}^{-1}\mathbf{g}(s_1) + 2(\mathbf{g}(s_2) - \mathbf{g}(s_1))^T\Sigma_{\boldsymbol{\eta}}^{-1}\mathbf{u}(s,t_0).$$

   Thus, the most probable stimulus given the observed time sample is found using the linear discriminant projection $\mathbf{w} = \Sigma_{\boldsymbol{\eta}}^{-1}(\mathbf{g}(s_2) - \mathbf{g}(s_1))$.

2. What is the ideal discrimination rule, for input measured in an interval $[0,T]$, assuming that the noise at different times is independent?

   Solution The optimal solution is to perform a time-averaged LDA by taking $\mathbf{w}\cdot\langle\mathbf{u}(s,t)\rangle_{t\in[0,T]}$ and comparing to the same threshold c, where $\langle\cdot\rangle_{t\in\mathcal{T}}$ is the sample mean over the set of time samples.

   This result follows directly from the single time sample case and the fact that $\log p(s_i|\mathbf{u}(s,t), t \in [0,T]) = \sum_{t=0}^{T}\log p(s_i|\mathbf{u}(t,s))$ for statistically independent samples.

Now, instead of an ideal observer, consider the information available after transformation of the noisy feedforward inputs by a recurrent population with rate $\mathbf{r} = [r_i]_{i=1}^N$. Let the population dynamics be given by

$$\tau_i\frac{\partial r_i(t,s)}{\partial t} = -r_i(t,s) + \phi_i\left(\sum_j W_{ij}r_j(t,s) + u_i(s,t)\right) \tag{4}$$

where τ_i are time constants, ϕ_i response nonlinearities and W_{ij} the connection weights (u_i is an element of input $\mathbf{u}$ defined above).

We are interested in $\mathcal{I}(s)$, the Fisher Information (FI) available about stimulus value s in the steady-state activity of this network $\mathbf{r}(s)$. We will approximate this by linearising the network around the noise-free limit, which (given Gaussian noise in $\mathbf{u}$) will result in Gaussian-distributed steady-state activity.

3. Setting $\boldsymbol{\eta}(t) = 0$ (so $\mathbf{u}(s,t) = \mathbf{g}(s)$), find the fixed point of the noise-free network activity $\mathbf{r}_0(s)$.

   Solution Set $\frac{\partial \mathbf{r}}{\partial t} = 0$ to give

$$\mathbf{r}_0(s) = \boldsymbol{\phi}\left(W\mathbf{r}_0(s) + \mathbf{g}(s)\right)$$

4. Show that the derivative of $\mathbf{r}_0$ with respect to s can be written

$$\mathbf{r}_0'(s) = -J^{-1}(s)\Phi'(s)\mathbf{g}'(s)$$

   where $J(s) = \Phi'(s)W - T^{-1}$, $T = \mathsf{diag}[\tau_i]$, and $\Phi'(s) = T^{-1}\mathsf{diag}\left[\phi_j'(\sum_k W_{jk}r_{0k}(s) + g_j(s))\right]$

   Solution Straightforward differentiation, with T cancelling in the final expression.

5. Now show that the network of eq. 4 can be linearised about $\mathbf{r}_0(s)$ (for small $\mathbf{r} - \mathbf{r}_0$ and small $\boldsymbol{\eta}$) to give

$$\frac{\partial \mathbf{r}(s,t)}{\partial t} = J(\mathbf{r}(s,t) - \mathbf{r}_0(s)) + \Phi'\boldsymbol{\eta}(t)$$

   Solution Linearising for $\mathbf{r} \approx \mathbf{r}_0$, we have:

$$T\frac{\partial \mathbf{r}(s,t)}{\partial t} \approx -\mathbf{r}_0 + \boldsymbol{\phi}(W\mathbf{r}_0 + \mathbf{u}) + (I + \Phi'W)(\mathbf{r} - \mathbf{r}_0)\,.$$

   Then linearise the $\boldsymbol{\phi}$ term for small $\boldsymbol{\eta}$ giving

$$T\frac{\partial \mathbf{r}(s,t)}{\partial t} \approx -\mathbf{r}_0 + \boldsymbol{\phi}(W\mathbf{r}_0 + \mathbf{g}) + T\Phi'\boldsymbol{\eta} + (I + T\Phi'W)(\mathbf{r} - \mathbf{r}_0)\,.$$

   Premultiplying by T and noting the value of $\mathbf{r}_0$ gives the result.

This first-order linearised system has the general solution:

$$\mathbf{r}(s,t) - \mathbf{r}_0(s) = e^{J(t-t_0)}(\mathbf{r}(s,t_0) - \mathbf{r}_0(s)) + \int_{t_0}^{t} e^{J(t-\tau)}\Phi'\boldsymbol{\eta}(\tau)d\tau$$

Provided the fixed point is stable (i.e., all eigenvalues of J have negative real part) we can take the stationary state limit by letting $t_0 \to -\infty$ to get:

$$\mathbf{r}(s,t) - \mathbf{r}_0(s) = \int_{-\infty}^{t} e^{J(t-\tau)}\Phi'\boldsymbol{\eta}(\tau)d\tau \tag{5}$$

Let the eigendecomposition of J be $J = V\Lambda V^{-1} = \sum_{i=1}^{N} \mathbf{v}_i^R \left(\mathbf{v}_i^L\right)^T \lambda_i$. Superscripts L and R denote left and right eigenvectors, which are the rows of V^{-1} and columns of V respectively.

6. Show that stationary-state covariance $\Sigma_{SS} = \left\langle (\mathbf{r}(s,t) - \mathbf{r}_0(s))(\mathbf{r}(s,t) - \mathbf{r}_0(s))^{\mathsf{T}} \right\rangle$ is given by

$$\Sigma_{SS} = -\sum_{i,j} \mathbf{v}_i^R \left(\mathbf{v}_i^L\right)^T \Phi'\Sigma_{\boldsymbol{\eta}}\Phi'\mathbf{v}_j^L \left(\mathbf{v}_j^R\right)^T \frac{1}{\lambda_i + \lambda_j}$$

Solution

$$\begin{aligned}\Sigma_{SS} &= \int_{-\infty}^{t}\int_{-\infty}^{t} e^{J(t-\tau)}\Phi'\Sigma_{\boldsymbol{\eta}}\delta(\tau-\tau')\Phi' e^{J^T(t-\tau')}d\tau\, d\tau' \\ &= \sum_{i,j}\mathbf{v}_i^R\left(\mathbf{v}_i^L\right)^T\Phi'\Sigma_{\boldsymbol{\eta}}\Phi'\mathbf{v}_j^L\left(\mathbf{v}_j^R\right)^T\int_{-\infty}^{t} e^{(\lambda_i+\lambda_j)(t-\tau)}d\tau \\ &= -\sum_{i,j}\mathbf{v}_i^R\left(\mathbf{v}_i^L\right)^T\Phi'\Sigma_{\boldsymbol{\eta}}\Phi'\mathbf{v}_j^L\left(\mathbf{v}_j^R\right)^T\frac{1}{\lambda_i+\lambda_j}\end{aligned}$$

Now make a simplifying change of variables $\tilde{\mathbf{r}} \equiv \Phi'^{-1}\mathbf{r}$ and $\tilde{J} \equiv \Phi'^{-1}J\Phi'$. In this basis, the linearised network becomes $\frac{\partial}{\partial t}\tilde{\mathbf{r}}(s,t) = \tilde{J}(\tilde{\mathbf{r}}(s,t)-\tilde{\mathbf{r}}_0(s))+\boldsymbol{\eta}(t)$ and $\tilde{J}$ has eigenvalues $\tilde{\lambda}_i = \lambda_i$ and eigenvectors $\tilde{\mathbf{v}}_i^L = \Phi'\mathbf{v}_i^L$, $\tilde{\mathbf{v}}_i^R = \Phi'^{-1}\mathbf{v}_i^R$.

7. Show that the linear FI (i.e. the Gaussian FI neglecting the dependence of Σ_{SS} on s) conveyed by the instantaneous network rate $\mathbf{r}(s,t)$ about s in the linearised model can be written in the form

$$\mathcal{I}_{lin} = \mathbf{g}'^T\left[\sum_{i,j}\tilde{\mathbf{v}}_i^R\left(\tilde{\mathbf{v}}_j^R\right)^T\Gamma_{ij}\right]^{-1}\mathbf{g}'.$$

   Give the form of Γ_{ij}.

   Solution We have $\mathcal{I}_{lin} = \boldsymbol{\mu}'^T\Sigma^{-1}\boldsymbol{\mu}'$. The mean of $\mathbf{r}$ under linearised dynamics is $\mathbf{r}_0$, and its covariance Σ_{SS}. So

$$\mathcal{I} \equiv \mathbf{r}_0'\cdot\Sigma_{SS}^{-1}\mathbf{r}_0' = -\mathbf{g}'^T\left[(\Phi')^{-1}\sum_{i,j}\mathbf{v}_i^R\left(\mathbf{v}_i^L\right)^{\mathsf{T}}\Phi'\Sigma_\eta\Phi'\mathbf{v}_j^L\left(\mathbf{v}_j^R\right)^T(\Phi')^{-1}\frac{\lambda_i\lambda_j}{\lambda_i+\lambda_j}\right]^{-1}\mathbf{g}'$$

   and we have

$$\Gamma_{ij} = -\left(\tilde{\mathbf{v}}_i^L\right)^T\Sigma_{\boldsymbol{\eta}}\tilde{\mathbf{v}}_j^L\frac{\lambda_i\lambda_j}{\lambda_i+\lambda_j} = \left(\tilde{V}^{-1}\Sigma_{\boldsymbol{\eta}}\tilde{V}^{-T}\right)_{ij}\frac{1}{\tau_i+\tau_j}$$

8. Relating this result to the ideal observer, describe the network properties likely to provide good discrimination.

   Solution Align slow modes to integrate $\Sigma_{\boldsymbol{\eta}}^{-1}\mathbf{g}'$.

4 Learning

In perceptual learning paradigms, subjects perform fine judgments on perceptual stimuli over extended periods of practice. One classic paradigm is fine orientation discrimination, in which subjects see oriented gratings rotated clockwise or counterclockwise relative to a reference angle and their task is to report the direction of rotation. Performance improves over many trials, sometimes over years. Here we will investigate a simple model of this setting using the deep learning framework.

Architecture: A three layer linear neural network consisting of an input layer with two units representing the angle of the presented gratings; a hidden layer with units representing V1 orientation-tuned neurons; and a decision or readout layer with a single unit encoding the decision made by the network. The network's computations are controlled by two weight matrices that encode the strength of synaptic interconnections between neurons. The first, $W^1 \in R^{N\times 2}$, encodes synaptic connectivity from the input layer to the N neurons in the V1-like representation layer, and the second, $W^2 \in R^{1\times N}$, encodes synaptic connectivity from the representation layer to the decision readout layer (a single output neuron). The network's output $\hat{y}$ in response to an input x is simply $\hat{y} = W^2W^1x$.

Task: The network is trained to produce a value of $y = +c$ in response to the input pattern $x = [1\ 0]^T$ and $y = -c$ in response to $x = [0\ 1]^T$, where the constant c controls the strength of the required response before it is deemed correct. These two inputs represent presentation of clockwise and counter-clockwise gratings respectively, and each occurs with probability 1/2 on any given trial.

Initialization: When a subject comes to this task, their primary visual cortex will already be orientation tuned. However, they won't know how to read out these orientations to do the task. We therefore will assume that V1 initially has bell curve tuning to orientation, while the readout layer starts at zero. In particular, let g_+ be a column vector of size N containing the population response to the clockwise orientation used during training, and g_- be a column vector containing the population response to the counterclockwise orientation. we assume that the population activity is relatively homogenous, that is, $||g_+||_2 = ||g_-||_2$. The initial weights of the representation layer are $W^1(0) = [g_+\ g_-] \in R^{N\times 2}$, to imbue it with orientation selectivity. The vector g_+ is thus the population response to the clockwise input, and g_- the response to the counterclockwise input. By contrast, the weights of the decision stage begin untuned, such that $W^2(0) = 0$.

Loss: Task performance is measuerd as the mean squared error between the desired output and the network's current output in response to an input x, that is, $\frac{1}{2}\langle(y - W^2W^1x)^2\rangle$.

Learning rule: To learn to perform the task, the weights W^1 and W^2 are changed over time in proportion to the negative gradient of the task performance. In the limit of small learning rates, such that learning is driven by the average statistics of both clockwise and counter-clockwise inputs, the gradient descent dynamics are

$$\tau\frac{d}{dt}W^1 = {W^2}^T(\Sigma^{yx} - W^2W^1\Sigma^x) \quad (1)$$

$$\tau\frac{d}{dt}W^2 = (\Sigma^{yx} - W^2W^1\Sigma^x){W^1}^T \quad (2)$$

where τ is a time constant inversely proportional to the learning rate, $\Sigma^{yx} = \langle xy^T \rangle$ is the input-output correlation matrix and $\Sigma^x = \langle xx^T \rangle$ is the input correlation matrix.

1. Compute Σ^{yx} and Σ^x for this task.

   (2 marks)

   Solution

   From the definition,

$$\begin{aligned} \Sigma^{yx} &= 1/2c \begin{bmatrix} 1 & 0 \end{bmatrix} - 1/2c \begin{bmatrix} 0 & 1 \end{bmatrix} && (3) \\ &= 1/2 \begin{bmatrix} c & -c \end{bmatrix} && (4) \\ \Sigma^x &= 1/2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 1/2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} && (5) \\ &= 1/2 I && (6) \end{aligned}$$

These dynamics are high dimensional, nonlinear, coupled, and hard to interpret. Let's find a reduction to two scalars. Let $d = g_+ - g_-$ be the difference in neural population responses to the two training orientations. Intuitively, it is this difference which must be uncovered and amplified during learning. Using this, we shall try to express the weights over time as

$$\begin{aligned} W^1(t) &= [g_+ \; g_-] + \alpha(t)[d \; -d], && (7) \\ W^2(t) &= \beta(t) d^T, && (8) \end{aligned}$$

where here $\alpha(t)$ and $\beta(t)$ are scalars that encode the amplification of the difference in population responses in the representation and decision layer weights respectively.

2. Derive the following reduction of the dynamics by substituting this form of the weights into the gradient descent dynamics:

$$\begin{aligned} \tau \frac{d}{dt} \alpha &= \frac{1}{2} \beta \left[c - v\beta(1 + 2\alpha) \right] && (9) \\ \tau \frac{d}{dt} \beta &= \frac{1}{2} \left[c - v\beta(1 + 2\alpha) \right] (1 + 2\alpha) && (10) \end{aligned}$$

   Here the constant $v = d^T g_+$ is something like a notion of task difficulty.

   (12 marks)

   Solution The parameter $v = d^T g_+ = g_+^T g_+ - g_+^T g_-$ measures the overlap between the difference vector and the input response, a metric of task difficulty. Because $||g_+|| = ||g_-||$, we further

have that $v = -d^T g_-$ and $d^T d = 2v$.

$$\begin{aligned}
\tau \frac{d}{dt} W^1 &= W^{2T}(\Sigma^{yx} - W^2 W^1 \Sigma^x), && (11)\\
\tau \frac{d}{dt} \alpha[d \ -d] &= \frac{1}{2}\beta d \left([c \ -c] - \beta d^T([g_+ \ g_-] + \alpha[d \ -d])\right), && (12)\\
&= \frac{1}{2}\beta d \left([c \ -c] - \beta([v \ -v] + \alpha[2v \ -2v])\right), && (13)\\
&= \frac{1}{2}\beta(c - v\beta(1+2\alpha))[d \ -d], && (14)\\
\tau \frac{d}{dt}\alpha &= \frac{1}{2}\beta(c - v\beta(1+2\alpha)). && (15)
\end{aligned}$$

And for the decision layer weights, we have

$$\begin{aligned}
\tau \frac{d}{dt} W^2 &= (\Sigma^{yx} - W^2 W^1 \Sigma^x) W^{1T}, && (16)\\
\tau \frac{d}{dt} \beta d^T &= \frac{1}{2}\left([c \ -c] - \beta d^T([g_+ \ g_-] + \alpha[d \ -d])\right)([g_+ \ g_-] + \alpha[d \ -d])^T, && (17)\\
&= \frac{1}{2}\left([c \ -c] - \beta([v \ -v] + \alpha[2v \ -2v])\right)([g_+ \ g_-] + \alpha[d \ -d])^T, && (18)\\
&= \frac{1}{2}\left(c - v\beta - 2v\alpha\beta\right)[1 \ -1]([g_+ \ g_-] + \alpha[d \ -d])^T, && (19)\\
&= \frac{1}{2}\left(c - v\beta(1+2\alpha)\right)(1+2\alpha)d^T, && (20)\\
\tau \frac{d}{dt}\beta &= \frac{1}{2}\left(c - v\beta(1+2\alpha)\right)(1+2\alpha). && (21)
\end{aligned}$$

A key question in the experimental literature is which layer changes most over learning. On one intuition, the representation layer should change the most, because it starts with better orientation tuning and so probably sets the maximum resolution possible. On another intuition, the readout layer should change the most, because the representation layer is already orientation tuned and so an appropriate readout can yield good performance.

3. We can answer this question without knowing the exact trajectory through time. Divide the differential equations from the reduction to obtain a differential equation describing how β evolves against α (that is, $d\beta/d\alpha = \cdots$). Solve this differential equation for the given initial conditions.

   (8 marks)

   Solution

   Dividing the differential equations, we have

$$\frac{d\alpha}{d\beta} = \frac{\beta}{1+2\alpha} \qquad (22)$$

which is separable with solution

$$\alpha + \alpha^2 = \beta^2/2 + C \tag{23}$$

where the constant C depends on the initial conditions. The relevant initial condition is no change in either the input or decision layer, $\alpha(0) = \beta(0) = 0$, and for this we have $C = 0$. Thus in terms of α, we have

$$\beta = \sqrt{2\alpha + 2\alpha^2} \tag{24}$$

4. Interpret this result. At the start of learning, which grows faster, α or β? If learning continues for a long time and α and β grow large, which is larger and by how much? Which intuition does gradient descent align with?

   (3 marks)

   Solution

   Initially, both weights start at zero. For small α, we therefore have $\beta \approx \sqrt{2}\sqrt{\alpha}$. Hence β grows more quickly than α when both are small. The decision layer will be the *first* to show large changes due to learning; it precedes α in time.

   Once α is large, then $\beta \approx \sqrt{2}\alpha$. This result shows that β eventually grows linearly with α, but is a factor of $\sqrt{2}$ larger in size. Hence the magnitude of the change in the decision layer after learning is *greater* than that in the input layer.

   Gradient descent therefore aligns with the idea that readout weights should change more, because the representation layer is already usefully orientation tuned.

So far we have investigated what happens in a network with just a single hidden layer. Now we will generalize this result to a simple deeper setting. Consider a deep linear chain with just a single neuron per layer, and scalar weights $a_1, \cdots, a^D$. The network output is $\hat{y} = a_D \cdots a_2 a_1 x$, where all quantities are scalars. We train this network again on the squared loss on a dataset with $\Sigma^{yx} = s$ and $\Sigma^x = \lambda$, both scalars.

The gradient dynamics on the weights are

$$\tau \frac{d}{dt} a_i = (s - \prod_{j=1}^{D} a_j \lambda) \prod_{j=1, j\neq i}^{D} a_j. \tag{25}$$

5. Compute $\tau \frac{d}{dt}(a_i^2)$ and show that the dynamics of the squared weights for all i are identical.

   (2 marks)

   Solution

   The dynamics are

$$\begin{aligned} \tau \frac{d}{dt}(a_i^2) &= 2(s - \prod_{j=1}^{D} a_j \lambda) \prod_{j=1}^{D} a_j && (26) \\ &= 2(s - u\lambda)u && (27) \\ &= f(u) && (28) \end{aligned}$$

where $u = a^D \cdots a_2 a_1$. Crucially, $f(u)$ does not depend on i. Hence all squared layer strengths change by the same amount, with dynamics driven only by the current overall product of weights.

6. Let $c_i \equiv a_i(t) - a_i(0)$ be the change in weight i from its initialization. Show that, if $a_i(0) < a_j(0)$ for some i, j then $c_i(t) > c_j(t)$ for all t.

   (10 marks)

   Solution

   At time t, each squared layer strength will have changed by the same factor $\Delta(t) = \int_0^t f(u)dt$. While it is difficult to compute $\Delta(t)$ as a function of time, the crucial point is that the change is the same for all the a_i's. Hence

   $$a_i(t)^2 = a_i(0)^2 + \Delta(t), \quad i = 1, \cdots, D, \tag{29}$$

   or without loss of generality, assuming positive a_i and $\Delta(t)$,

   $$a_i(t) = \sqrt{a_i(0)^2 + \Delta(t)}. \tag{30}$$

   The change in strength from the beginning of learning is thus

   $$c_i(t) \equiv a_i(t) - a_i(0) = \sqrt{a_i(0)^2 + \Delta(t)} - a_i(0). \tag{31}$$

   Suppose we have $a_i(0) < a_j(0)$ for some i, j. Then we claim $c_i(t) > c_j(t), \ \forall t$.

   To prove this we show that the function $g(x) = \sqrt{x^2 + c} - x$ is monotone decreasing for positive $c > 0$. Its derivative is

   $$\frac{d}{dx} g(x) = \frac{x}{\sqrt{x^2 + c}} - 1. \tag{32}$$

   To show this is always less than zero we use proof by contradiction,

   $$\frac{x}{\sqrt{x^2 + c}} - 1 \geq 0 \tag{33}$$

   $$x \geq \sqrt{x^2 + c} \tag{34}$$

   $$0 \geq c \tag{35}$$

   which contradicts the assumption $c > 0$. Hence since $c_i(t) = g(a_i(0))$ and $c_j(t) = g(a_j(0))$ with $c = \Delta(t)$, $a_i(0) < a_j(0)$ implies $c_i(t) > c_j(t)$.

7. Interpret this result. Which layers change most in a deep chain?

   (3 marks)

   Solution

   This means that the weakest layers change the most. If we have some order of initial weight strengths $a_i(0) < a_j(0) < \cdots$, then the size of the changes will be the reverse order, $c_i(0) > c_j(0) > \cdots$. In a deep linear chain, which layer changes most is not about proximity to the output, but about the initialization.