Gatsby Computational Neuroscience Unit
Theoretical Neuroscience

Final examination, theoretical neuroscience
13 May 2024

Part I – short questions

There are four sections with four questions each. Please answer three out of each four, starting the answers for each section on a new page. Don't forget to write your name at the top of each block of answers.

Good luck!

1 Biophysics

1. **Build your own realistic model of release probability.** Vesicles are divided more or less into two groups (it's really three, but we'll keep it simple): a readily releasable pool and a reserve pool. Vesicles in the readily releasable pool can be released when an action potential arrives at the presynaptic terminal, but they have to be replenished from the reserve pool, which happens at a relatively slow rate (100s of ms). Write down stochastic equation for the release probability as a function of presynaptic spike times. Your model should be sufficiently detailed that you could simulate it on a computer.

   Solution

   What matters for release is the number of vesicles in the readily releasable pool (RRP). So we want to write down a stochastic update rule for that number, which we'll denoted k. The update rule depends on the release probability and replenishment rate, which we'll denote $P_{\rm rel}(k)$ and $\nu(k)$, respectively. Their explicit definitions, and properties, are

   - $P_{\rm rel}(k)$: the release probability given k vesicles in the RRP. $P_{\rm rel}(k)$ is a decreasing function of k, and it's zero when $k = 0$.
   - $\nu(k)$: the rate at which vesicles move from the reserve pool to the RRP. $\nu(k)$ is an increasing function of k, and it's zero when $k = k_{\rm max}$.

   Given these definitions, we have

   - If there's no spike at time t, then

   $$k(t+dt) = \begin{cases} k(t) & \text{probability } 1-\nu(k)dt \\ k(t)+1 & \text{probability } \nu(k)dt \end{cases} \tag{1}$$

   - If there is a spike at time t, then

   $$k(t+dt) = \begin{cases} k(t) & \text{probability } 1-P_{\rm rel}(k) \\ k(t)-1 & \text{probability } P_{\rm rel}(k)\,. \end{cases} \tag{2}$$

   Note that with probability $\nu(k)dt$, k can increase, but in the small dt limit that can be ignored.

   We want to write this as an update rule for the $P(k,t)$, the probability that there are k vesicles in the readily releasable pool. Again we have two conditions,

   - If there's no spike at time t, then

   $$P(k,t+dt) = \big(1-\nu(k)dt\big)P(k,t) + \nu(k-1)dtP(k-1,t)\,, \tag{3}$$

   which implies that

   $$\frac{dP(k,t)}{dt} = \nu(k-1)P(k-1,t) - \nu(k)P(k,t)\,. \tag{4}$$

   - If there is a spike at time t, then

   $$P(k,t+dt) = \big(1-P_{\rm rel}(k)\big)P(k,t) + P_{\rm rel}(k+1)P(k+1,t)\,, \tag{5}$$

   which implies that

   $$\frac{dP(k,t)}{dt} = \sum_i \delta(t-t_i)\big(P_{\rm rel}(k+1)P(k+1,t) - P_{\rm rel}(k)P(k,t)\big) \tag{6}$$

   where t_i is the time of the $i^{\rm th}$ spike.

   In these expressions, it's understood that, for all time, $P(-1,t) = P(k_{\rm max}+1,t) = 0$.

2. **Axons with different diameters.** Consider an infinitely long axon in which the radius is a_1 for $x < 0$ and a_2 for $x > 0$. You inject constant current at $x = 0$, so $I = I_0 L\delta(x)$ (the factor of L is a length scale; it's needed to make the units come out). Write down the steady state voltage as a function of x (with x both positive and negative). Make sure to work out the amplitude as a function of $I_0 L$.

Solution

We want to solve the cable equation in steady state

$$\lambda(x)^2 \frac{\partial^2 V(x)}{\partial x^2} = V(x) - r_M I_0 L \delta(x) \tag{7}$$

where λ, the electrotonic length, is given in general by

$$\lambda^2(x) = \frac{r_M a(x)}{2 r_L} \,. \tag{8}$$

Fortunately, the x-dependence of the radius, a, is pretty simple, so we can solve the cable equation separately for $x < 0$ and $x > 0$. That gives us

$$V(x) = V_0 \left(e^{x/\lambda_1} \Theta(-x) + e^{-x/\lambda_2} \Theta(x) \right) \tag{9}$$

where $\lambda_1 \equiv r_M a_1 / 2 r_L$ and $\lambda_2 \equiv r_M a_2 / 2 r_L$, and $\Theta(x)$ is the Heaviside step functions (it's 1 if $x > 0$ and 0 if $x < 0$). This solution matches one boundary condition: the voltage is continuous at $x = 0$. However, we have to match the other boundary condition: the current is continuous. Since current is injected at $x = 0$, $V(x)$ must be discontinuous at $x = 0$. To determine how discontinuous, we take two derivatives. The first derivative is given by

$$\lambda(x)^2 \frac{\partial V(x)}{\partial x} = V_0 \left(\lambda_1 e^{x/\lambda_1} \Theta(-x) - \lambda_2 e^{-x/\lambda_2} \Theta(x) \right) . \tag{10}$$

This is not continuous: the voltage goes from $V_0 \lambda_1$ when x is slightly smaller than 0 to $-V_0 \lambda_2$ when x is slightly greater than zero. The derivative of this function at $x = 0$ is, then, given by

$$\lambda(x)^2 \frac{\partial^2 V(x)}{\partial^2 x} \bigg|_{x=0} = -V_0 (\lambda_1 + \lambda_2) \delta(x) \,. \tag{11}$$

Comparing this to Eq. (7), we see that V_0 is given by

$$V_0 = r_M I_0 \frac{L}{\lambda_1 + \lambda_2} \,. \tag{12}$$

3. **A somewhat strange Hodgin-Huxley motivated neuron.** Consider the following neuron,

$$\tau \frac{dV}{dt} = -(V - \mathcal{E}_L) - g_{Na} m_\infty(V)(V - \mathcal{E}_{Na}) - g_K n (V - \mathcal{E}_k) \tag{13a}$$

$$\tau_n \frac{dn}{dt} = m_\infty(V) - n \,. \tag{13b}$$

That's not a typo: we're assuming m and n have the same activation function, which has the usual form,

$$m_\infty(V) = \frac{1}{1 + \exp(-(V - V_m)/\Delta V)} \,. \tag{14}$$

Use the following parameters,

$$\tau = 10 \text{ ms} \tag{15a}$$

$$\tau_n = 10 \text{ ms} \tag{15b}$$

$$\mathcal{E}_L = -65 \text{ mV} \tag{15c}$$

$$\mathcal{E}_{Na} = +20 \text{ mV} \tag{15d}$$

$$\mathcal{E}_K = -80 \text{ mV} \tag{15e}$$

$$V_m = -50 \text{ mV} \tag{15f}$$

$$\Delta V = 5 \text{ mV} \tag{15g}$$

$$g_{Na} = 30 \tag{15h}$$

$$g_K = 170 \,. \tag{15i}$$

(a) Show that there's a stable fixed point at $V = \mathcal{E}_L$.

(b) Assume that the voltage has been at rest for a long time (meaning long compared to all relevant time points), and suddenly jumps to -30 mV. Sketch, qualitatively, voltage versus time from that point onward.

Hint: you could draw nullclines, but qualitative arguments plus linear stability analysis should be enough. I chose parameters so that the fixed point is especially easy to find!

Solution

(a) First we need to find the steady state. For n that's particularly easy: $dn/dt = 0$ implies $n = m_\infty(V)$. Inserting that into the equation for voltage, we find that

$$\tau \frac{dV}{dt} = -(V - \mathcal{E}_L)\big(1 + (g_{Na} + g_K) m_\infty(V)\big)\,. \tag{16}$$

Thus, the steady state membrane potential is $V = \mathcal{E}_L$. Linearizing around that fixed point and keeping track only of the signs, it's easy to show that

$$\begin{pmatrix} \tau \frac{d\delta V}{dt} \\ \tau_n \frac{d\delta n}{dt} \end{pmatrix} = \begin{pmatrix} - & - \\ + & - \end{pmatrix} \begin{pmatrix} \delta V \\ \delta n \end{pmatrix} \tag{17}$$

Given the signs, the trace is negative and the determinant is positive, and so both eigenvalues are negative.

(b) Because this is a 2-D system and there's only one fixed point, we know the voltage will eventually return at $\mathcal{E}_L$. To determine its trajectory, note that when $V \to$ -30 mV, dV/dt immediately increases, and the voltage rapidly rises. Because $\tau_n = 10$ ms, it takes a little time for the potassium channel to open and pull the voltage down. Thus, you'll get a spike with a width of a few ms and a height probably somewhere between about -20 and 0 mV, followed by relaxation to rest, at -65 mV.

4. **Type II neurons with slow adaptation.** Consider a reduced model of a neuron,

$$\tau \frac{dr}{dt} = -ar + 2r^2 - r^3 \tag{18a}$$

$$\tau\epsilon \frac{d\theta}{dt} = 1 - (1 - \epsilon^2)\cos\theta \tag{18b}$$

$$\tau_a \frac{da}{dt} = r - \frac{1}{2} \tag{18c}$$

with $\tau_a \gg \tau \gg \epsilon$ and $0 < \epsilon < 1$. The membrane potential, V, is given by

$$V = V_0 + V_1 r \sin^2(\theta/2)\,. \tag{19}$$

The θ-equation is chosen so actual spikes emerge (in the small ϵ limit), but it's not part of the question.

Show that this model exhibits bursting: alternating periods of firing ($r > 0$) and quiescence ($r = 0$), with the time of each proportional to τ_a.

Solution

The right hand side of Eq. (18a) has fixed points at

$$r = 0, 1 \pm \sqrt{1 - a}\,. \tag{20}$$

First consider $a > 0$. In that case, the fixed point at $r = 0$ is stable, and if $a < 1$, there are two more fixed points. Of those, the larger one is stable and the smaller one is unstable (as is easy to see just by plotting the right hand side of Eq. (18a)). Thus, in the range $0 < a < 1$, there are stable fixed points at $a = 0$ and $a > 1$. When a exceeds 1, the two fixed points with $r > 0$ vanish, leaving a single fixed point at $r = 0$.

If, on the other hand, $a < 0$, the fixed point at zero is unstable. Now, according to Eq. (20) the stable fixed points are at $a < 0$ and $a > 0$.

Let's say we start with $a > 0$ and $r = 0$. In that case, Eq. (18c) tells us that a decreases. When a falls below zero, r increases until it reaches the single positive fixed point. According to Eq. (20), that fixed point is at $a > 1$. Consequently, according to Eq. (18c), a increases again. When it exceeds 1, the fixed points at $r > 0$ vanish, and r decreases to zero. Where the process starts over again.

For this clean picture to emerge, we need a to change slowly, which will happen if τ_a is large enough. In that case, the characteristic timescale for bursting is the time it takes a to change, which is proportional to τ_a.

2 Networks

1. Consider a classical Hopfield network, whose update rule is

$$S_i(t+1) = \text{sign}\left[\frac{1}{N}\sum_{j\neq i}^{N}\left(\sum_{\nu=1}^{P}\xi_i^\nu\xi_j^\nu\right)S_j(t)\right] \tag{1}$$

where the ξ_i^ν are pulled *iid* from a random binary vectors,

$$\xi_i^\nu = \begin{cases} +1 & \text{probability } 1/2 \\ -1 & \text{probability } 1/2\,. \end{cases} \tag{2}$$

At time $t = 0$, the network is initialized according to

$$S_i(0) = \begin{cases} \xi_i^\mu & \text{probability } q \\ 1 & \text{probability } 1-q\,. \end{cases} \tag{3}$$

What is the probability that $S_i(t) = \xi_i^\mu$? Your answer will depend on N and P as well as q.

Solution

Separating the terms with $\nu = \mu$ and $\nu \neq \mu$, Eq. (1) becomes

$$S_i(1) = \text{sign}\left[\xi_i^\mu\frac{1}{N}\sum_j\xi_j^\mu S_j(0) + \sum_{\nu\neq\mu}\xi_i^\nu\xi_j^\nu\frac{1}{N}\sum_j S_i(0)\right]. \tag{4}$$

Note first of all that

$$\xi_j^\mu S_j(0) = \begin{cases} 1 & \text{probability } q \\ \xi_j^\mu & \text{probability } 1-q\,. \end{cases} \tag{5}$$

Since $\xi_j^\mu = \pm 1$ with equal probability, this means

$$\xi_j^\mu S_j(0) = \begin{cases} +1 & \text{probability } (1+q)/2 \\ -1 & \text{probability } (1-q)/2\,. \end{cases} \tag{6}$$

Using this, and ignoring $1/N$ corrections associated with $j \neq i$, the average of the first sum in Eq. (4) is

$$\frac{1}{N}\sum_j\langle\xi_j^\mu S_j(0)\rangle = \frac{1+q}{2} - \frac{1-q}{2} = q \tag{7}$$

And, again ignoring $1/N$ corrections, it's variance is

$$\text{Var}\left[\frac{1}{N}\sum_j\xi_j^\mu S_j(0)\right] = \frac{1}{N^2}\sum_j\left\langle\left(\xi_j^\mu S_j(0) - q\right)^2\right\rangle = \frac{1}{N}\left(\frac{1+q}{2}(1-q)^2 + \frac{1-q}{2}(1+q)^2\right) = \frac{1-q^2}{N}\,. \tag{8}$$

For the sum over ν and j, we note that the product $\xi_i^\nu\xi_j^\nu S_0(0)$ is uncorrelated across i, j and ν, and zero mean. Thus, the sum is Gaussian (with respect to index, i) with variance $(N-1)(P-1)$. Taking into account the division by N, we thus have

$$\sum_{\nu\neq\mu}\xi_i^\nu\xi_j^\nu\frac{1}{N}\sum_j S_i(0) \sim \left(\frac{(N-1)(P-1)}{N^2}\right)^{1/2}\eta_i \approx \left(\frac{P}{N}\right)^{1/2}\eta_i \tag{9}$$

where η_i is a zero mean, unit variance Gaussian random variable.

Putting all this together, and using the central limit theorem (which tells us that large sums are Gaussian), we arrive at

$$S_i(1) = \text{sign}\left[q\xi_i^\mu + \sigma\zeta_i\right] \tag{10}$$

where ζ_i is a zero mean, unit variance Gaussian random variable and

$$\sigma^2 \equiv \frac{P+1-q^2}{N} \approx \frac{P}{N}\,. \tag{11}$$

So we really could have ignored the first variance. Thus, the probability that $S_i(1) = \xi_i^\mu$ is the probability that $\sigma\zeta_i > -q$. More formally,

$$P\big(S_i(1) = \xi_i^\mu\big) = H\left(\frac{q}{\sigma}\right) \tag{12}$$

where H is the cumulative normal function.

2. Consider a network of excitatory and inhibitory neurons which have a stable fixed point given by

$$\nu_E = \phi_E\big(J_{EE}\,\nu_E - J_{EI}\nu_I + h_E\big) \tag{13a}$$
$$\nu_I = \phi_I\big(J_{IE}\,\nu_E - J_{II}\nu_I + h_I\big) \tag{13b}$$

where all weights are positive. Let $h_E \to h_E + \delta h_E$ where δ_E is infinitesimally small and positive. In the limit of large weights, do the excitatory and inhibitory firing rates increase or decrease?

Solution

In the limit of large weights, we have

$$J_{EE}\delta\nu_E - J_{EI}\delta\nu_I + \delta h_E = 0 \tag{14a}$$
$$J_{IE}\delta\nu_E - J_{II}\delta\nu_I = 0\,. \tag{14b}$$

Which implies that

$$\delta\nu_E = \frac{J_{II}\delta h_E}{D} \tag{15a}$$
$$\delta\nu_I = \frac{J_{IE}\delta h_E}{D} \tag{15b}$$

where

$$D \equiv J_{EI}J_{IE} - J_{EE}J_{II}\,. \tag{16}$$

For stability, the determinant of the linearized dynamics must be positive. In the limit of large weights, D is proportional to the determinant. Thus, both ν_E and ν_I increase.

3. Consider a network whose equilibrium is given by

$$x_i = \phi\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{n} w_{ij}x_j\right)\,. \tag{17}$$

As usual, you want to approximate the sum as a Gaussian random variable. However, unlike as usual, the weights are correlated, and they have the following statistics,

$$\langle w_{ij}\rangle = 0 \tag{18a}$$
$$\langle w_{ij}w_{kl}\rangle = \sigma^2\left(\delta_{ik}\delta_{jl} + \frac{\rho}{N}(1 - \delta_{ik}\delta_{jl})\right)\,. \tag{18b}$$

Let

$$y_i = \frac{1}{N^{1/2}}\sum_{j=1}^{n} w_{ij}x_j \tag{19}$$

and define

$$\mathrm{Var}[y] \equiv \frac{1}{N}\sum_i \left(\frac{1}{N^{1/2}}\sum_{j=1}^{n} w_{ij}x_j\right)^2\,. \tag{20}$$

Assuming N is large, write an expression for $\mathrm{Var}[y]$ in terms of σ^2, ρ and the first and second moments of x, defined to be

$$\overline{x} = \frac{1}{N}\sum_j x_j \tag{21a}$$
$$\overline{x^2} = \frac{1}{N}\sum_j x_j^2\,. \tag{21b}$$

Solution

The variance of y can be written

$$\mathrm{Var}[y] = \frac{1}{N^2}\sum_{ijj'} w_{ij}w_{ij'}x_j x_{j'} = \frac{1}{N}\sum_{jj'} x_j x_{j'} \frac{1}{N}\sum_i w_{ij}w_{ij'} \,. \tag{22}$$

Consider first $j = j'$, for which

$$\frac{1}{N}\sum_i w_{ij}^2 \approx \sigma^2 \tag{23}$$

where the approximation gets increasingly better as N increases. If instead $j \neq j'$, we have

$$\frac{1}{N}\sum_i w_{ij}w_{ij'} \approx \frac{\rho_w \sigma^2}{N} \,. \tag{24}$$

Putting these together gives us

$$\frac{1}{N}\sum_i w_{ij}w_{ij'} \approx \sigma^2\left(\delta_{ij} + \frac{\rho_w}{N}(1-\delta_{ij})\right) \approx \sigma^2\left(\delta_{ij} + \frac{\rho_w}{N}\right) \,. \tag{25}$$

Inserting this into Eq. (22), we arrive at

$$\mathrm{Var}[y] \approx \sigma^2\left(\overline{x^2} + \rho_w \overline{x}^2\right) \,. \tag{26}$$

4. Consider a convolutional neuronal network in one dimension with Mexican-hat connectivity:

$$\tau\frac{\partial v(x,t)}{\partial t} = -v(x,t) + \int W(x-x')\phi(v(x',t))dx' + I(x,t) \tag{27}$$

Where ϕ is a monotonic input-output function, $I(x)$ is an external input and $W(x-x')$ is the connection strength between units at x and x', given by a difference of Gaussians:

$$W(x) = \frac{1}{\sigma_2 - \sigma_1}\left(\sigma_2 e^{-\frac{x^2}{2\sigma_1^2}} - \sigma_1 e^{-\frac{x^2}{2\sigma_2^2}}\right) \qquad \text{with } \sigma_1 < \sigma_2 \,. \tag{28}$$

Show that the critical spatial frequency that first emerges when the homogeneous fixed-point solution to a constant input ($I(x,t) = I_0$) loses stability (i.e. spontaneous pattern formation) is:

$$k^* = \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sigma_2 - \sigma_1}\left[\left(\frac{\sigma_1}{\sigma_2}\right)^{\frac{\sigma_1^2}{\sigma_2^2-\sigma_1^2}} - \left(\frac{\sigma_1}{\sigma_2}\right)^{\frac{\sigma_2^2}{\sigma_2^2-\sigma_1^2}}\right] \,. \tag{29}$$

Reminder:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} e^{\pm ikx} dx = e^{-\frac{k^2\sigma^2}{2}} \tag{30}$$

Solution First one needs to compute the homogeneous steady-state solution to a constant input. If the input is constant we assume a homogeneous solution $v(x) = v_0$:

$$v_0 = \int W(x-x')\phi(v_0)dx' + I_0 = I_0 \tag{31}$$

Where we used that $\int W(y)dy = 0$. We then need to check the stability of the steady state. For that we make a perturbation $v(x,t) = v_0 + \delta v(x,t)$ and derive under which conditions the system relaxes back to v_0.

$$\tau \frac{\partial \delta v(x,t)}{\partial t} = -v_0 - \delta v(x,t) + \int_{-\infty}^{\infty} W(x-x')\phi(v_0 + \delta v(x',t))dx' + I_0 \tag{32a}$$

$$\tau \frac{\partial \delta v(x,t)}{\partial t} = -\delta v(x,t) + \phi'(v_0) \int_{-\infty}^{\infty} W(x-x')\delta v(x',t)dx' \,. \tag{32b}$$

Because the system is linear we propose a plane wave solution $\delta v(x,t) = c(t)e^{-ikx}$

$$\tau \dot{c}(t)e^{-ikx} = -c(t)e^{-ikx} + \phi'(v_0)c(t) \int_{-\infty}^{\infty} W(x-x')e^{-ikx'}dx' \tag{33a}$$

$$\tau \dot{c}(t) = -c(t) + \phi'(v_0)c(t) \int_{-\infty}^{\infty} W(y)e^{iky}dy \tag{33b}$$

$$\tau \dot{c}(t) = -c(t)(1 - \phi'(v_0)\hat{W}(k)) \,. \tag{33c}$$

Which means that the perturbation will decay exponentially if

$$1/\phi'(v_0) > \hat{W}(k) \,. \tag{34}$$

We therefore need to find the upper bound of

$$\hat{W}(k) = \frac{\sqrt{2\pi}\sigma_2\sigma_1}{\sigma_2 - \sigma_1} \left(e^{-\frac{\sigma_1^2 k^2}{2}} - e^{-\frac{k^2\sigma_2^2}{2}} \right) \tag{35a}$$

Taking the derivative equal to zero, and asking $k \neq 0$ (that is the minimum), we find k^*

$$\sigma_1^2 k^* e^{-\frac{\sigma_1^2 k^{*2}}{2}} - k^*\sigma_2^2 e^{-\frac{k^{*2}\sigma_2^2}{2}} = 0 \tag{36a}$$

$$\frac{\sigma_2^2}{\sigma_1^2} = e^{\frac{k^{*2}}{2}(\sigma_2^2 - \sigma_1^2)} \tag{36b}$$

$$k^{*2} = \frac{2}{\sigma_2^2 - \sigma_1^2} \ln \frac{\sigma_2^2}{\sigma_1^2} \tag{36c}$$

Replacing in the expression for $\hat{W}(k)$

$$\hat{W}(k^*) = \frac{\sqrt{2\pi}\sigma_2\sigma_1}{\sigma_2 - \sigma_1} \left(e^{-\frac{\sigma_1^2}{\sigma_2^2 - \sigma_1^2} \ln \frac{\sigma_2^2}{\sigma_1^2}} - e^{-\frac{\sigma_2^2}{\sigma_2^2 - \sigma_1^2} \ln \frac{\sigma_2^2}{\sigma_1^2}} \right) \tag{37}$$

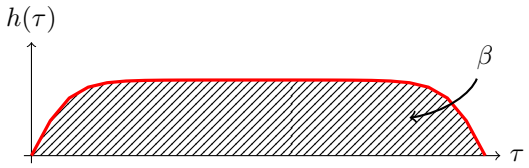
Which is equivalent to Eq. (29).

3 Coding

1. A model of the spiking neuronal response to stimulus $x(t)$ is given by the generalised linear intensity function

$$\lambda(t|H(t)) = e^{\alpha x(t) - \int_0^\infty h(\tau)s(t-\tau)d\tau}\,.$$

where $h(\tau) \geq 0$ with $\int_0^\infty h(\tau)d\tau = \beta$ and $s(t)$ is the δ-function representation of the spike train. Let $h(\tau)$ be a smooth function, roughly constant over a long support.



   (a) Describe the qualitative behaviour you expect from such a model (note the minus sign in the definition of λ!).

   (b) Now suppose that the stimulus is held fixed over the duration of a trial, so that the firing of the neuron comes to a steady-state mean rate. Using a mean-field approximation (i.e. exchanging an expectation and exponential), find an expression constraining the steady-state rate. Use this to find the derivative of the rate with respect to the sustained stimulus value x.

   (c) Show that, in this limit, the neuron behaves similarly to a soft-threshold activation function rate neuron.

   Solution

   (a) The model will exhibit spike-rate adaptation, but otherwise generate Poisson-like spiking.

   (b) Let the steady-state mean intensity function be $\bar{\lambda}(x)$. Making the mean-field assumption we have

$$\bar{\lambda} = e^{\alpha x - \int_0^\infty d\tau h(\tau)\bar{\lambda}} = e^{\alpha x - \beta\bar{\lambda}}$$

   Differentiating with respect to x we have

$$\begin{aligned}
\frac{d\bar{\lambda}}{dx} &= \left(\alpha - \beta\frac{d\bar{\lambda}}{dx}\right) e^{\alpha x - \beta\bar{\lambda}} \\
&= \left(\alpha - \beta\frac{d\bar{\lambda}}{dx}\right)\bar{\lambda} \\
(1+\beta\bar{\lambda})\frac{d\bar{\lambda}}{dx} &= \alpha\bar{\lambda} \\
\frac{d\bar{\lambda}}{dx} &= \frac{\alpha}{\frac{1}{\bar{\lambda}} + \beta}
\end{aligned}$$

   (c) We see that for small $\bar{\lambda}$ the derivative approaches $\alpha\bar{\lambda}$ corresponding to an exponential rise. For large $\bar{\lambda}$ the derivative approaches α/β and so the rate grows linearly.

2. A population of neurons encodes an angular value θ with von Mises tuning curves

$$f_i(\theta) = \alpha_i e^{\kappa_i cos(\theta - \theta_i)}\,.$$

If the neurons generate spikes with independent Poisson statistics given these means, show that the posterior estimate of θ (asuming a uniform prior) is in a von-Mises-like exponential family. What is the base measure?

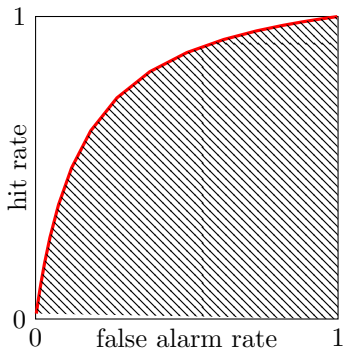
Solution

This is very easy to finf using the canonical exponential family form, with $\boldsymbol{\theta}$ a unit vector with angle θ:

$$\begin{aligned}
P(\mathbf{n}|\theta) &= \prod_i \frac{1}{n_i!} f_i(\theta)^{n_i} e^{f_i(\theta)} \\
&\propto \prod_i e^{n_i \mathbf{k}_i^\intercal \boldsymbol{\theta}} e^{f_i(\theta)} \\
&= e^{(\sum_i n_i \mathbf{k}_i)^\intercal \boldsymbol{\theta} + \sum_i f_i(\theta)}
\end{aligned}$$

So the base measure is proportional to $e^{\sum_i f_i(\theta)}$.

3. Recall that the signal detection theory model of behaviour in a two-alternative choice assumes that the stimulus is mapped to a one-dimensional noisy *decision variable* x, and that choices are based on comparing this variable to a threshold value. As the threshold increase, the hit- and false-alarm-rates will both decrease monotonically. The relationship between them is represented by the receiver operating characteristic (ROC) curve.



Show that the area under the ROC curve corresponds to the probabilty that, for independent draws x_0 and x_1 of x given a true negative and true positive input respectively, $x_1 > x_0$.

Solution

Consider a threshold t. Then the ROC curve for threshold t represents the point $(C_0(t), C_1(t))$, where $C_i(t) = P(x_i > t)$ is the CDF of x_i.

The area under the curve is

$$\begin{aligned} A &= \int_{-\infty}^{\infty} C_1(t) dC_0(t) \\ &= \int_{-\infty}^{\infty} C_1(t) p_0(t) dt \\ &= \int_{-\infty}^{\infty} p(x_1 > t \wedge x_0 = t) dt \\ &= P(x_1 > x_0) \end{aligned}$$

where the third line required independence.

4. Suppose that the (standard, i.e. "right") eigenvectors of a stable D-dimensional linear dynamical system described by matrix A are all close to the same $(D-1)$-dimensional subspace, but otherwise random. That is, there is a $D \times D-1$ matrix P with orthonormal columns such that, for each unit-norm $\mathbf{v}_i : A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ $(i = 1 \dots D)$

$$\|PP^{\mathsf{T}}\mathbf{v}_i\| > 1 - \epsilon$$

for a small positive ϵ.

Contrast the generic response of the system to a unit-norm input either (a) orthogonal to; or (b) within the subspace defined by (the columns of) P. Your answer may be qualitative, but make your reasoning clear.

Solution

Let V be the matrix of normalised eigenvectors. Then an input $\mathbf{b}$ will load onto the eigenmodes according to $V^{-1}\mathbf{b}$. Let $\tilde{\mathbf{v}}_i$ be the ith row of V^{-1} viewed as a vector (the normalised $\tilde{\mathbf{v}}$ are sometimes called the "left eigenvectors" of A). These loadings decay (since A is stable) multiplicatively according to the propagator $e^{\lambda_i t}$ (or λ_i^t in discrete time).

Now, the fact that $V^{-1}V = I$ implies that $\tilde{\mathbf{v}}_i^{\mathsf{T}}\mathbf{v}_i = 1$ and $\tilde{\mathbf{v}}_i^{\mathsf{T}}\mathbf{v}_j = 0$ for $j \neq i$. The second condition means that every $\tilde{\mathbf{v}}_i$ must be oriented almost orthogonal to the space of P. This in turn makes it almost orthogonal to $\mathbf{v}_i$, and so the first condition implies that it must have a large norm.

(a) An input orthogonal to P will be nearly aligned with the $\tilde{\mathbf{v}}_i$, and so in the generic case the magnitude of the loading onto each eigenmode will be given by their (large) norms. Initially, these large loadings cancel in the output, thus reconstructing the unit-length $\mathbf{b}$. However, as faster-decaying modes drop away, the persistently large loadings on slower modes will be revealed, leading to transient amplification. Eventually, activity will be dictated by the slowest mode and will rotate to lie within P.

(b) An input within P will be almost orthogonal to all the $\tilde{\mathbf{v}}_i$, and so the loading onto each eigenmodes will generically have roughly unit strength. Thus, while the pattern of decay will be the same, we do not expect any non-normal transient amplification.

4 Learning

1. Suppose we train a linear student network on a dataset drawn from a linear teacher network. The teacher produces data $y = \bar{w}X + \epsilon$ where $y \in R^{1\times P}$ are the target outputs, $\bar{w} \in R^{1\times N}$ are the teacher weights, and $X \in R^{N\times P}$ is a matrix of P inputs drawn i.i.d. from an N-dimensional Gaussian with mean zero and variance $1/N$, $X_{ij} \sim \mathcal{N}(0, 1/N)$. Here $\epsilon \in R^{1\times P}$ is noise added to the teacher for each example, drawn i.i.d. from a zero mean Gaussian with variance σ_e^2. The teacher weights are drawn i.i.d. from a zero mean Gaussian with variance $\sigma_{\bar{w}}^2$. A key parameter of the problem is the signal-to-noise ratio $\mathcal{S} = \frac{\sigma_{\bar{w}}^2}{\sigma_e^2}$. Rather than train the student with gradient descent, we will take a more abstract view and suppose that we choose the student weights $w \in R^{1\times N}$ according to

$$w^* = \text{argmin}_w \frac{1}{2}||y - wX||_2^2 + \frac{\gamma}{2}||w||_2^2$$

where γ is a regularisation parameter. In the high dimensional regime where $P, N \to \infty$ with load $\alpha = P/N$, the normalised generalisation error can be shown to be

$$\frac{E_g}{\sigma_{\bar{w}}} = \int \rho(\lambda) \left[\frac{\gamma^2}{(\lambda+\gamma)^2} + \frac{\lambda}{\mathcal{S}(\lambda+\gamma)^2} \right] d\lambda + \frac{1}{\mathcal{S}}$$

where $\rho(\lambda)$ is the Marcenko-Pastur distribution.

   (a) What is the optimal strength of regularisation to use?

   (b) Recall that the unregularised setting yields the double descent phenomenon in which intermediate amounts of data ($\alpha = 1$) exhibit the worst generalization error. Reasoning intuitively, would you expect the optimal regularisation to exhibit double descent as α grows?

   (c) Suppose now that the teacher is noiseless, but we are in a partially observable setting in which the student can only see a fraction ρ of the total features in the teacher. That is, the teacher has N weights as usual, but the student has only $\approx \rho N$ weights that connect to a fixed subset of the input vector. What is the optimal regularisation in terms of ρ?

   (d) A potential advantage of being a partially observable student is that such a student contains fewer parameters compared to the fully observed one. Reasoning intuitively, should the student ever choose to ignore inputs?

Solution

(a) We differentiate and set the result equal to zero.

$$\begin{aligned}
0 &= \frac{\partial}{\partial\gamma} \int \rho(\lambda) \left[\frac{\gamma^2}{(\lambda+\gamma)^2} + \frac{\lambda}{\mathcal{S}(\lambda+\gamma)^2} \right] d\lambda && (1)\\
0 &= \int \rho(\lambda) \left[2\gamma(\lambda+\gamma)^{-2} - 2\gamma^2(\lambda+\gamma)^{-3} - 2\frac{\lambda}{\mathcal{S}}(\lambda+\gamma)^{-3} \right] d\lambda && (2)\\
0 &= 2\gamma(\lambda+\gamma)^{-2} - 2\gamma^2(\lambda+\gamma)^{-3} - 2\frac{\lambda}{\mathcal{S}}(\lambda+\gamma)^{-3} && (3)\\
0 &= \gamma(\lambda+\gamma) - \gamma^2 - \frac{\lambda}{\mathcal{S}} && (4)\\
0 &= \gamma\lambda - \frac{\lambda}{\mathcal{S}} && (5)\\
\gamma &= \frac{1}{\mathcal{S}} && (6)
\end{aligned}$$

(b) There is no double descent in this case–the regularisation suppresses very small eigenvalues, which are the source of the problem.

(c) From the student's perspective, the unviewed part of the teacher is equivalent to label noise. Let $\tilde{w} \in R^{1\times(1-\rho)N}$ be the teacher weights to the inputs that are not observed by the student, and $\tilde{x} \in R^{(1-\rho)N\times 1}$ be the vector of those inputs. The contribution to the output on any given example is $\tilde{\epsilon} = \tilde{w}\tilde{x}$, so

$$\begin{aligned}
\tilde{\epsilon} &\sim \mathcal{N}(0, \tilde{w}\frac{1}{N} I_{(1-\rho)N} \tilde{w}^\top) && (7)\\
&\sim \mathcal{N}(0, \sigma_{\bar{w}}^2 \frac{(1-\rho)N}{N}) && (8)\\
&\sim \mathcal{N}(0, \sigma_{\bar{w}}^2(1-\rho)). && (9)
\end{aligned}$$

A teacher with noise variance $\sigma_e^2 = \sigma_w^2(1-\rho)$ has signal-to-noise ratio $\mathcal{S} = \frac{\sigma_w^2}{\sigma_w^2(1-\rho)} = \frac{1}{1-\rho}$. Because the optimal regularisation level depends only on SNR and not on α, the change in dimensionality of the student is irrelevant. The optimal regularisation is therefore $\gamma = 1-\rho$.

(d) For the noise free teacher, double descent is not an issue. Ignoring inputs adds noise, potentially causing a double descent peak without proper regularisation.

2. Consider a strange neural network that receives a single scalar x as input and computes its output as $\hat{y} = (w_2^2 + w_1^2)x$, where w_2 and w_1 are scalar weights. We measure the performance of the network using some per-example loss $l(y, \hat{y})$ averaged over a dataset, yielding the total loss

$$\mathcal{L} = \langle l(y, \hat{y}) \rangle .$$

The weights are updated using gradient flow,

$$\tau \dot{w}_i = -\frac{\partial \mathcal{L}}{w_i} \quad \text{for } i = 1, 2. \tag{10}$$

(a) Show that the network's output is invariant to rotating the weights (that is, consider the two weights as a vector and apply the rotation matrix $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ to obtain new weights rotated by an angle θ).

(b) What point is equal to itself after being rotated by any angle?

(c) Show that this point is a critical point of the gradient dynamics.

Solution

(a) Rotating the weights by an angle θ yields new weights

$$\begin{bmatrix} w_1' \\ w_2' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \tag{11}$$

$$= \begin{bmatrix} w_1\cos\theta - w_2\sin\theta \\ w_1\sin\theta + w_2\cos\theta \end{bmatrix}, \tag{12}$$

and input-output function

$$[(w_1')^2 + (w_2')^2]x = [(w_1\cos\theta - w_2\sin\theta)^2 + (w_1\sin\theta + w_2\cos\theta)^2]x \tag{13}$$

$$= [w_1^2\cos^2\theta - w_1w_2\cos\theta\sin\theta + w_2^2\sin^2\theta \tag{14}$$

$$+ w_1^2\sin^2\theta + w_1w_2\cos\theta\sin\theta + w_2^2\cos^2\theta]x \tag{15}$$

$$= [w_1^2 + w_2^2]x. \tag{16}$$

(b) The point $w_1 = w_2 = 0$ is always equal to itself when rotated.

(c) Consider a point a distance ϵ away from the origin. We can rotate this point while leaving the network function unchanged, tracing out a circle. Regardless of the loss, the gradient must be perpendicular to the tangent of this circle. Because we can make this circle arbitrarily close to the origin, we see that at the origin, the gradient must be orthogonal to every direction–which can happen only if the gradient is zero. Hence the origin is a critical point.

3. Two linear neurons with weights $w_1, w_2 \in R^{1\times N}$ and input $x \in R^{N\times 1}$ are trained with Oja's rule

$$\dot{w}_i = \eta y_i(x^\top - y_i w_i) \quad \text{for } i = 1, 2, \tag{17}$$

where η is a learning rate. However, we also add inhibition from neuron 1 to neuron 2, such that

$$y_1 = w_1 x \tag{18}$$

$$y_2 = w_2(x - w_1^\top y_1). \tag{19}$$

(a) What will the weights of neuron 1 typically converge to?

(b) What will the weights of neuron 2 typically converge to?

(c) How would you generalise this scheme to more neurons and what algorithm does it implement?

(d) Biologically, is this a reasonable scheme?

Solution

(a) For neuron 1, we have

$$\begin{aligned}
\langle \dot{w}_1 \rangle &= \eta \left\langle y_1 (x^\top - y_1 w_1) \right\rangle && (20)\\
&= \eta \left\langle w_1 x (x^\top - w_1 x w_1) \right\rangle && (21)\\
&= \eta \left\langle w_1 x x^\top - w_1 x x^\top w_1^\top w_1 \right\rangle && (22)\\
&= \eta w_1 C - \eta w_1 C w_1^\top w_1. && (23)
\end{aligned}$$

Finding fixed points by setting the derivative to zero we have

$$\begin{aligned}
0 &= \eta w_1 C - \eta w_1 C w_1^\top w_1 && (24)\\
0 &= w_1 C - w_1 C w_1^\top w_1 && (25)\\
w_1 C &= w_1 C w_1^\top w_1 && (26)
\end{aligned}$$

such that w_1 is a left eigenvector of C or zero. We know that from random initializations we will tend to get alignment to the principle eigenvector $w_1(t) \to \alpha e_1^\top$. The proportionality constant must satisfy

$$\begin{aligned}
0 &= \eta w_1 C - \eta w_1 C w_1^\top w_1 && (27)\\
0 &= \alpha e_1^\top e_1 \lambda e_1^\top - \alpha^3 e_1^\top e_1 \lambda e_1^\top e_1 e_1^\top && (28)\\
0 &= \alpha \lambda e_1^\top - \alpha^3 \lambda e_1^\top && (29)\\
0 &= \alpha (1 - \alpha^2) && (30)
\end{aligned}$$

and so $\alpha = 1$ at the non-zero fixed point. Hence the first neuron's weights converge to $e_1^\top$, the principle eigenvector of the input correlation matrix.

(b) For neuron 2, eventually the first neuron will have converged to $e_1^\top$. We then have the following dynamics

$$\begin{aligned}
\langle \dot{w}_2 \rangle &= \eta \left\langle y_2 (x^\top - y_2 w_2) \right\rangle && (31)\\
&= \eta \left\langle w_2 (x - w_1^\top y_1)(x^\top - w_2 (x - w_1^\top y_1) w_2) \right\rangle && (32)\\
&= \eta \left\langle w_2 (x - e_1 e_1^\top x)(x^\top - w_2 (x - e_1 e_1^\top x) w_2) \right\rangle && (33)\\
&= \eta w_2 (I - e_1 e_1^\top) C - \eta w_2 (I - e_1 e_1^\top) C (I - e_1 e_1^\top)^\top w_2^\top w_2. && (34)\\
& && (35)
\end{aligned}$$

From this, we see that the dynamics are driven by a new correlation matrix in which the maximal eigenvector direction has been projected to zero. Hence by a similar line of reasoning as before, we know $w_2(t) \to e_2$.

(c) To generalise this scheme, each neuron j can receive inhibition from all neurons preceding it, such that its output is

$$y_j = w_j \left(x - \sum_{k<j} w_k^\top y_k \right). \tag{36}$$

It would compute PCA of the input data.

(d) While it's possible that some part of the brain could perform an operation like this, it isn't likely. It requires precise connectivity such that the inhibition onto the input of other neurons is the transpose of the feed forward input. Furthermore, it implies that some neurons would be substantially more active (the higher variance PCA modes) than others, which we don't see empirically.

4. The brain is not a strict feedforward network. Instead, there are often skip connections that shortcut layers of the network. Suppose we have a deep linear network with skip connections, that produces its output as

$$\hat{y} = w_2 w_1 x + w_s x \tag{37}$$

where $w_2, w_1 \in R$ are scalar feedforward weights and $w_s \in R$ is a scalar skip connection straight from input to output. Here for simplicity everything is scalar. We train this network to minimize the squared error averaged over a dataset $\mathcal{D}$,

$$\mathcal{L} = \left\langle \frac{1}{2}(y - \hat{y})^2 \right\rangle_{\mathcal{D}}. \tag{38}$$

Suppose the dataset has statistics $\Sigma^{yx} = \langle yx \rangle_{\mathcal{D}}$ and $\Sigma^x = \langle x^2 \rangle_{\mathcal{D}}$. The weights are updated via gradient flow,

$$\tau \dot{w}_i = -\frac{\partial \mathcal{L}}{\partial w_i} \quad \text{for } i = 1, 2, s. \tag{39}$$

(a) What are the gradient flow equations for each weight in terms of Σ^{yx} and Σ^x?

(b) Suppose we start with an initial condition such that $w_1(0) = w_2(0)$. Let $w_d = w_2 w_1$ be the total weight in the deep pathway. What is $\frac{d}{dt} w_d$ under the gradient flow dynamics?

(c) Show that $\frac{1}{2}\log(w_d) - w_s$ is a conserved quantity under these dynamics (that is, its time derivative is zero).

(d) Use this to argue that w_s is the dominant pathway when Σ^{yx}/Σ^x is small, and w_d is the dominant pathway when Σ^{yx}/Σ^x is very large, when the weights are initialized to a small constant value $w_d(0) = w_s(0) = \epsilon$.

Solution

(a) The chain rule yields

$$\begin{aligned}
\tau \dot{w}_2 &= \langle (y - w_2 w_1 x + w_s x) w_1 x \rangle_{\mathcal{D}} && (40)\\
&= (\langle yx \rangle_{\mathcal{D}} - (w_2 w_1 + w_s) \langle x^2 \rangle_{\mathcal{D}}) w_1 && (41)\\
&= (\Sigma^{yx} - (w_2 w_1 + w_s)\Sigma^x) w_1 && (42)\\
\tau \dot{w}_1 &= (\Sigma^{yx} - (w_2 w_1 + w_s)\Sigma^x) w_2 && (43)\\
\tau \dot{w}_s &= \Sigma^{yx} - (w_2 w_1 + w_s)\Sigma^x. && (44)
\end{aligned}$$

(b) When initialized with equal values, w_1 and w_2 remain equal under the dynamics. The time derivative is

$$\begin{aligned}
\tau \frac{d}{dt} w_2 w_1 &= w_2 \tau \dot{w}_1 + \tau \dot{w}_2 w_1 && (45)\\
&= (w_2^2 + w_1^2)(\Sigma^{yx} - (w_2 w_1 + w_s)\Sigma^x) && (46)\\
&= 2 w_2 w_1 (\Sigma^{yx} - (w_2 w_1 + w_s)\Sigma^x) && (47)\\
&= 2 w_d (\Sigma^{yx} - (w_d + w_s)\Sigma^x) && (48)
\end{aligned}$$

(c) We have

$$\begin{aligned}
\tau \frac{d}{dt} \frac{1}{2} \log(w_d) - w_s &= \frac{1}{2 w_d} \tau \dot{w}_d - \tau \dot{w}_s && (49)\\
&= \frac{1}{2 w_d} 2 w_d (\Sigma^{yx} - (w_d + w_s)\Sigma^x) - (\Sigma^{yx} - (w_d + w_s)\Sigma^x) && (50)\\
&= 0. && (51)
\end{aligned}$$

(d) We know that $\frac{1}{2}\log(w_d) - w_s = C$ for some C. Therefore

$$\begin{aligned}
\sqrt{w_d} &= e^{w_s + C} && (52)\\
w_d &= e^C e^{2 w_s}. && (53)
\end{aligned}$$

We also know the solution manifold is $w_d + w_s = \Sigma^{yx}/\Sigma^x$. Starting from small weights $w_d = w_s = \epsilon$, the log term will dominate and $C = \frac{1}{2}\log(\epsilon) - \epsilon$ will approach negative infinity. Therefore w_d will be small compared to w_s if the task does not demand large $w_d + w_s$. However if the problem requires the total weight to grow sufficiently, then w_d will grow exponentially compared to w_s, eventually surpassing it. Intuitively, the deep pathway initially makes slow progress compared to the shallow pathway, but once it gets started, it grows faster.