Gatsby Computational Neuroscience Unit
Theoretical Neuroscience

Final examination, theoretical neuroscience
15 May 2024

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

Consider a linear integrate and fire neuron augmented by the so-called H-current,

$$\tau \frac{dV}{dt} = -(V - \mathcal{E}_L) - x g_0 (V - \mathcal{E}_0) \tag{1a}$$

$$\tau_x \frac{dx}{dt} = x_\infty(V) - x \tag{1b}$$

where

$$x_\infty(V) = \frac{1}{1 + \exp\big((V - V_0)/\Delta V\big)} \,. \tag{2}$$

When the voltage reaches V_{th}, a spike is emitted and the voltage is instantaneously rest to $\mathcal{E}_L$. The numerical values of the parameters are

$$\mathcal{E}_L = -65 \text{ mV} \tag{3a}$$

$$V_{\text{th}} = -50 \text{ mV} \tag{3b}$$

$$\mathcal{E}_0 = 0 \text{ mV} \tag{3c}$$

$$V_0 = -70 \text{ mV} \tag{3d}$$

$$\Delta V = 10 \text{ mV} \tag{3e}$$

$$\tau = 10 \text{ ms} \tag{3f}$$

$$\tau_x = 1000 \text{ ms} \,. \tag{3g}$$

You'll be choosing the value of g_0.

1. Ignoring the fact that the neuron can spike, sketch the nullclines when $g_0 = 1$.

   (5 marks)

   Solution The x-nullcline is given by Eq. (2). For the V-nullcline, we set dV/dt to zero in Eq. (1a), which gives us the equation (using $\mathcal{E}_0 = 0$)

   $$V = \frac{\mathcal{E}_L}{1 + x g_0} \,, \tag{4}$$

   which we can invert to find x in terms of V,

   $$x = \frac{\mathcal{E}_L / V - 1}{g_0} \,, \tag{5}$$

   The resulting nullclines are drawn in Fig. 1.

2. You should have one fixed point. Is it stable or unstable? Justify your answer.

   (5 marks)

   Solution

   For this you can just look at the nullclines and draw trajectories. Or do formal stability analysis, which proceeds as follows.

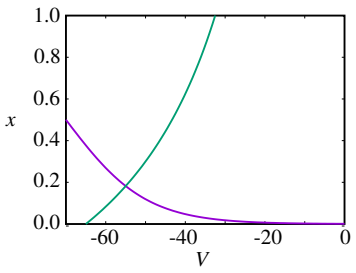


Figure 1: Nullclines. Green: V-nullcline; purple: x-nullcline.

The linearized dynamics is given by

$$\frac{d}{dt}\begin{pmatrix} V \\ x \end{pmatrix} = \begin{pmatrix} -\tau^{-1}(1+g_0 x) & -\tau^{-1} g_0 V \\ \tau_x^{-1} x'_\infty(V) & -\tau_x^{-1} \end{pmatrix} \begin{pmatrix} V \\ x \end{pmatrix} \tag{6}$$

For stability, we need the trace to be positive and determinant to be negative. The trace is clearly negative. The determinant, denoted D, is given by

$$D = (\tau\tau_x)^{-1}\big((1+xg_0) + g_0 V x'_\infty(V)\big) \tag{7}$$

where x and V need to be evaluated at the intersection of the nullclines. Because both V and $x'_\infty(V)$ are negative, the second term is negative, and so the determinant is positive. Thus, the fixed point is stable.

3. What are the conditions on g_0 that will guarantee repetitive firing?

   (10 marks)

   Solution

   For repetitive firing, the nullclines must cross at a value of the voltage that's above threshold. So we'll compute x on the two nullclines evaluated at $V = V_{\rm th} = -50$ mV,

$$x\text{-nullcline:}\quad x = \frac{1}{1+e^{(-50+70)/10}} = \frac{1}{1+e^2} \tag{8a}$$

$$V\text{-nullcline:}\quad x = \frac{(-65)/(-50)-1}{g_0} = \frac{0.3}{g_0}\,. \tag{8b}$$

   The condition for repetitive firing is, then,

$$\frac{1}{1+e^2} > \frac{0.3}{g_0}\,, \tag{9}$$

   which in turn implies that

$$g_0 > 0.3(1+e^2)\,. \tag{10}$$

4. Suppose you modify Eq. (1a) by adding external drive,

$$\tau\frac{dV}{dt} = -(V-\mathcal{E}_L) - xg_0(V-\mathcal{E}_0) + V_X\,. \tag{11}$$

Make a qualitative plot of firing rate versus V_X, including negative values (for V_x). Pick a reasonable value for G_0.

(10 marks)

Solution

Adding external drive shifts the V-nullcline, and causes the neuron to fire if the nullclines cross at a voltage above threshold. So we just need to calculate the new V-nullcline,

$$V = \frac{\mathcal{E}_L + V_X}{1 + x g_0}\,, \tag{12}$$

which we can invert to find x in terms of V,

$$x = \frac{(\mathcal{E}_L + V_x)/V - 1}{g_0}\,. \tag{13}$$

The difference in the value of x between the x-nullcline and V-nullclines evaluated at $V = V_{\rm th} = -50$ mV, denoted Δx, is given by

$$\begin{aligned} \Delta x &= \frac{1}{1+e^2} - \frac{(-65 + V_X)/-50 - 1}{g_0} \\ &= \frac{1}{1+e^2} - \frac{0.3 - V_X/50}{g_0} \\ &= \frac{1}{1+e^2} - \frac{0.3}{g_0} + \frac{V_X}{50 g_0}\,. \end{aligned} \tag{14}$$

The firing rate is zero when the right hand side is negative, and is an increasing function of V_X. Any plot with those characteristics will do.

5. Suppose you added a standard Hodgin-Huxley voltage-dependent potassium current to the neuron. Suppose also that g_0 is less than the value found in part 3 (which means that without the added potassium current, the neuron wouldn't fire on its own). Show that if the time constant of the potassium current is long enough, the neuron could exhibit two regimes: silent and repetitive firing.

   (10 marks)

   Solution

   We'll assume the potassium current isn't active at rest. In that case, because g_0 is less than the value found in part 3, the neuron can't fire on its own. However, if it does start firing, when it's hyperpolarized (remember that the potassium reversal potential is around -80 mV), x increases, which can cause the neuron to fire.

2 Networks

Consider a network in which the excitatory neurons are divided into M populations and the inhibitory population is homogeneous. Using ν_α to denote the average firing rate associated with excitatory population α, ν_I to denote the average firing rate of the inhibitory neurons, and ignoring quenched noise associated with variability in firing rate, a reasonable set of equations for each of the populations is

$$\tau_E \frac{d\nu_\alpha}{dt} = \phi\left(W_{EE}\nu_\alpha + cW_{EE}\sum_{\beta\neq\alpha}\nu_\beta - W_{EI}\nu_I + h'_\alpha\right) - \nu_\alpha \tag{1a}$$

$$\tau_I \frac{d\nu_I}{dt} = \phi\left(W_{IE}\sum_{\beta}\nu_\beta - W_{II}\nu_I + h'_I\right) - \nu_I \tag{1b}$$

where all constants (τ, W, c and h) are positive and $\phi(\cdot)$ is positive, bounded, continuous, and a monotonic increasing function of its argument. Make the following assumptions about the parameters: the weights and the h's are large and c is not (for instance, $W_{QR}, h'_Q \sim \mathcal{O}(\sqrt{K}\,)$ and $c \sim \mathcal{O}(1)$ with $K \to \infty$), and $\tau_I \ll \tau_E$.

1. Show that we can eliminate inhibition, and write a single equation for the excitatory populations,

$$\tau_E \frac{d\nu_\alpha}{dt} = \phi\left(W_0\nu_\alpha - W_1\sum_{\beta}\nu_\beta + h_\alpha\right) - \nu_\alpha\,. \tag{2}$$

where

$$W_0 \equiv W_{EE}(1-c) \tag{3a}$$

$$W_1 \equiv \frac{W_{EI}W_{IE}}{W_{II}} - cW_{EE} \tag{3b}$$

$$h_\alpha \equiv h'_\alpha - \frac{W_{EI}}{W_{II}}h'_I\,. \tag{3c}$$

(5 marks)

Solution

In the limit of fast inhibition ($\tau_I \ll \tau_E$), we can set the left hand side of Eq. (1b) to zero. That still gives us a nonlinear equation, but for large weights and h's, the term in parentheses must go to zero. That allows us to express ν_I in terms of the ν_α as

$$\nu_I = \frac{W_{IE}\sum_\beta \nu_\beta + h'_I}{W_{II}}\,. \tag{4}$$

Inserting that into Eq. (1a) and performing a small amount of algebra yields the desired result.

2. Show that when $h_\alpha > 0$ for all α and $0 < W_0 < W_1$, equilibria with more than one of the ν_α nonzero are unstable and equilibria with only one of the ν_α nonzero are stable. In other words, the stable equilibria (of which there can be M) all have $\nu_\alpha \neq 0$ and $\nu_\beta = 0$ for $\beta \neq \alpha$.

   (15 marks)

   Solution

   Because the weights are large, the linearized dynamics includes only the terms inside the parentheses in Eq. (2). In vector notation, this gives us

$$\frac{d\delta\boldsymbol{\nu}}{dt} \propto (W_0\mathbf{I} - W_1\mathbf{1}\mathbf{1}) \cdot \delta\boldsymbol{\nu} \tag{5}$$

   where $\delta\boldsymbol{\nu}$ is distance from the equilibrium firing rate, $\mathbf{I}$ is the identity, and $\mathbf{1}$ is a vector consisting of all 1's. For any vector $\delta\boldsymbol{\nu}$ such that $\mathbf{1} \cdot \delta\boldsymbol{\nu} = 0$, the eigenvalue is W_0, which is positive. However, once there's only one population left, there is no nonzero vector for which $\mathbf{1} \cdot \delta\boldsymbol{\nu} = 0$. In this case the only eigenvalue is $W_0 - W_1$, which is negative. Thus, that equilibrium is stable.

3. Derive conditions on the h_α that will allow stable equilibria with $\nu_\alpha > 0$ and $\nu_{\beta\neq\alpha} = 0$. The equilibrium should be stable for all α.

   (10 marks)

   Solution

   Let's say ν_α is nonzero and and $\nu_{\beta\neq\alpha} = 0$. Using Eq. (2), at equilibrium ν_α is given by

$$\nu_\alpha = \frac{h_\alpha}{W_1 - W_0}\,. \tag{6}$$

   For ν_β to have a stable equilibrium at $\nu_\beta = 0$, the relevant term in the parentheses in Eq. (2) must be negative – which, because the weights are large, will drive ν_β to zero. Setting ν_β to zero, the term inside the parentheses in Eq. (2) is negative when

$$-W_1\nu_\alpha + h_\beta < 0\,. \tag{7}$$

   Using the above expression for ν_α and setting ν_β to zero, that can be written

$$h_\beta < \frac{W_1 h_\alpha}{W_1 - W_0}\,. \tag{8}$$

   That must be true for all α, β pairs. Consequently, if we want M stable equilibria involving only one population each, Eq. (9) must be satisfied for all α, β pairs.

4. Draw the nullclines for a two population model ($M = 2$), assuming h_1 and h_2 satisfy

$$\frac{h_\alpha}{h_\beta} < \frac{W_1}{W_1 - W_0} \tag{9}$$

   for $\alpha = 1$ and $\beta = 2$, and also for $\alpha = 2$ and $\beta = 1$. Include arrows indicating the direction of flow on either side of the nullclines, and indicate which fixed points are stable and which are unstable.

(10 marks)

Solution

In the limit of large weights, the nullclines are given by (from Eq. (2))

$$\nu_1 \text{ nullcline:} \quad (W_0 - W_1)\nu_1 - W_1\nu_2 + h_1 = 0 \tag{10a}$$

$$\nu_2 \text{ nullcline:} \quad (W_0 - W_1)\nu_2 - W_1\nu_1 + h_2 = 0\,, \tag{10b}$$

which can we written

$$\nu_2 = -\rho\nu_1 + \frac{h_1}{W_1} \tag{11a}$$

$$\nu_2 = -\frac{1}{\rho}\nu_1 + \frac{h_2}{\rho W_1} \tag{11b}$$

where

$$\rho \equiv \frac{W_1 - W_0}{W_1}\,. \tag{12}$$

These nullclines are plotted in Fig. 2, with (to reduce clutter) W_1 set to 1. The intercepts are labeled on each axis. Equation (9) implies that both $h_1 < h_2/\rho$ and $h_2 < h_1/\rho$, so the nullclines do indeed cross.

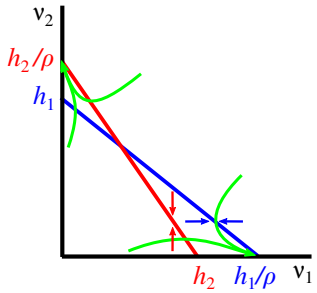


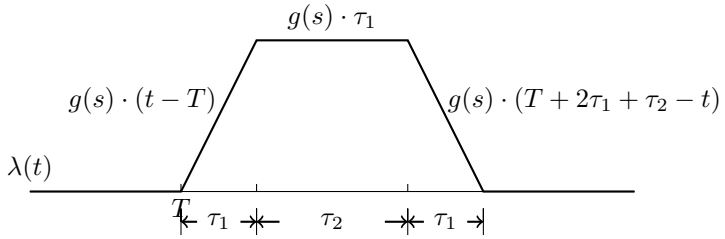
Figure 2: Nullclines. Blue: ν_1 nullcline; red: ν_2 nullcline. Blue and red arrows are directions of trajectories for ν_1 and ν_2, respectively. Green arrows are sample trajectories.

3 Coding

It has occasionally been suggested that, in addition to the intensity, the **latency** of sensory neuronal responses might encode stimulus parameter values. We will build a simple model of this idea.

Consider cells that are described by inhomogenous Poisson statistics. In the absence of any stimulus their intensity functions are all 0. After stimulus onset, the intensity functions increase linearly for a period τ_1, with a slope that depends on the stimulus parameter s according to a "latency tuning curve" $g(s)$. The firing rate then remains constant for a period τ_2, and then falls off linearly with slope $-g(s)$. Thus, if the stimulus onset is at time T, the intensity function of a single cell is given by

$$\lambda(t) = \begin{cases} 0 & t \in (0, T], \\ g(s) \cdot (t - T) & t \in (T, T + \tau_1], \\ g(s) \cdot \tau_1 & t \in (T + \tau_1, T + \tau_1 + \tau_2], \\ g(s) \cdot (T + 2\tau_1 + \tau_2 - t) & t \in (T + \tau_1 + \tau_2, T + 2\tau_1 + \tau_2] \end{cases}$$



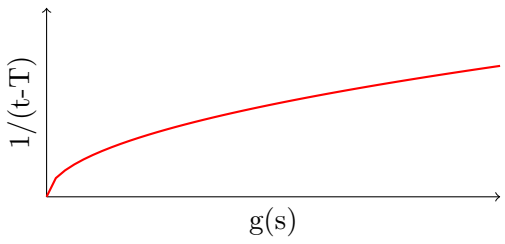
1. Calculate the density function, for repeated presentations of the same stimulus, for the time of the first spike evoked from such a cell (10 marks). Assume that the initial linear rise is long enough that you can neglect the possibility that the first spike occurs after the plateau is reached. Sketch the inverse of the mean first-spike latency as a function of $g(s)$ (2.5 marks).

   Solution

$$\begin{aligned} p(t)dt &\propto P\{N[T,t) = 0\}P\{N[t,t+dt) = 1\} \\ &= e^{-\int_T^t \lambda(\tau)d\tau}\lambda(t)dt \\ &= e^{-\int_T^t g(s)(\tau - T)d\tau}g(s)(t-T)dt \\ &= g(s)(t-T)e^{-\frac{1}{2}g(s)(t-T)^2}dt \end{aligned}$$

   This is Weibull in (t-T) with $\alpha = 2$, $\beta = \frac{1}{2}g(s)$. So the expectation is

$$\mathbb{E}[t - T \mid g(s)] = \left(\frac{g(s)}{2}\right)^{-\frac{1}{2}} \Gamma\left(1 + \frac{1}{2}\right)$$



2. How large is the Fisher information about s provided by the time of first spike from a single cell (15 marks)? How does your answer compare to the Fisher information obtained from the total number of spikes in the response (5 marks)?

   Solution

$$\begin{aligned}
FI_{\text{latency}} &= \left\langle -\frac{d^2}{ds^2}\log p(t-T|g(s)) \right\rangle \\
&= \left\langle -\frac{d^2}{ds^2}\left(\log g(s) + \log(t-T) - \frac{1}{2}g(s)(t-T)^2\right) \right\rangle \\
&= \left\langle -\left(\frac{g''(s)}{g(s)} - \frac{g'(s)^2}{g(s)^2} - \frac{1}{2}g''(s)(t-T)^2\right) \right\rangle \\
&= \frac{g'(s)^2}{g(s)^2} - \frac{g''(s)}{g(s)} + \frac{1}{2}g''(s)\left\langle (t-T)^2 \right\rangle \\
&= \frac{g'(s)^2}{g(s)^2} - \frac{g''(s)}{g(s)} + \frac{1}{2}g''(s)\frac{2}{g(s)}\Gamma(2) \\
&= \frac{g'(s)^2}{g(s)^2}
\end{aligned}$$

   The total number of spikes is Poisson, with mean $\mu(s) = g(s)\tau_1(\tau_2+\tau_1)$. Using the expression for Poisson FI from lecture we have

$$\begin{aligned}
FI_{\text{count}} &= \frac{\mu'(s)^2}{\mu(s)} \\
&= \frac{g'(s)^2}{g(s)}\tau_1(\tau_2+\tau_1)
\end{aligned}$$

   so the latency FI is larger provided

$$g(s) < \frac{1}{\tau_1(\tau_2+\tau_1)}\,.$$

3. Latency codes are often criticised on the basis that decoding requires precise knowledge of the time of stimulus onset. Explain how a population code (of cell with responses as described above) can avoid this problem. You may assume that although the exact time of stimulus onset is not known to the decoder, stimuli are spaced far enough apart to identify which spike is first after a stimulus. (7.5 marks).

Solution

Each first spike induces a joint distribution over T and s. Labelling the neurons with i, and assuming a uniform prior for T:

$$p(s, T|t_i) \propto g_i(s)(t_i - T)e^{-\frac{1}{2}g_i(s)(t_i-T)^2}p(s)$$

and assuming independence

$$p(s, T|\mathbf{t}) \propto \prod_i g_i(s)(t_i - T)e^{-\frac{1}{2}\sum_i g_i(s)(t_i-T)^2}p(s)$$

With enough neurons, this will concentrate allowing identification of both s and T.

It might help to know that the nth ($n = 1, 2, \ldots$) non-central moment for the Weibull distribution with density

$$p(t) = \alpha\beta\, t^{\alpha-1}e^{-\beta t^\alpha}$$

is

$$E[t^n] = \beta^{-n/\alpha}\Gamma\left(1 + \frac{n}{\alpha}\right).$$

4 Learning

Deep neural networks can learn representations suitable for solving nonlinear tasks. Here we investigate their learning dynamics on a simple class of tasks. We will qualitatively understand the dynamics of training a single hidden layer ReLU network by reducing it to several deep linear networks using a gated deep linear network formalism.

Dataset The canonical example of a nonlinear problem is the XoR problem. We consider a slight generalization: an XoR problem in the first two input dimensions, but a linearly separable problem in the third. Our dataset $X \in R^{3\times 4}, y \in R^{1\times 4}$ consists of four data points in three dimensions,

$$X = \begin{bmatrix} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ -\Delta & \Delta & \Delta & -\Delta \end{bmatrix} \tag{1}$$

$$y = \begin{bmatrix} -1 & 1 & 1 & -1 \end{bmatrix} \tag{2}$$

where $\Delta \geq 0$ is a parameter that controls the linear separability of the points along the third dimension.

Architecture We will try to solve this task using a ReLU network producing output

$$\hat{y} = W_2 \max(0, W_1 x),$$

with weights $W_2 \in R^{N_o \times N_h}$, $W_1 \in R^{N_h \times N_i}$ and N_o, N_h, where N_i are the output, hidden, and input dimensions respectively.

Loss function The network performance is measured as the mean squared error over the dataset

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

Learning Rule Each weight matrix is updated via gradient flow

$$\dot{W}_i = -\frac{\partial \mathcal{L}}{\partial W_i} \quad \text{for } i = 1, 2.$$

Our goal will be to understand how the network transitions from a regime in which the data points are clearly linearly separable ($\Delta >> 1$) to the regime in which they are not ($\Delta = 0$, the classical XOR task).

1. Suppose we have a ReLU network with a single hidden neuron ($N_h = 1$). Show that the dynamics are equivalent to a deep linear neural network's gradient flow dynamics on the effective dataset statistics $\Sigma^{yx} = \left\langle \theta(W_1 x) y x^\top \right\rangle_{\mathcal{D}}, \Sigma^x = \left\langle \theta(W_1 x) x x^\top \right\rangle_{\mathcal{D}}$, where $\theta(n)$ is the Heaviside function which is 0 for $n < 0$ and 1 otherwise.

   (10 marks)

Solution We have the gradient flow updates

$$\begin{aligned}
\dot{W}_1 &= \left\langle W_2 \left[y - W_2 \text{max}(0, W_1 x) \right] \theta(W_1 x) x^\top \right\rangle_{\mathcal{D}} && (3)\\
&= \left\langle W_2 \left[y - W_2 \theta(W_1 x) W_1 x \right] \theta(W_1 x) x^\top \right\rangle_{\mathcal{D}} && (4)\\
&= W_2 \left[\left\langle \theta(W_1 x) y x^\top \right\rangle_{\mathcal{D}} - W_2 W_1 \left\langle \theta(W_1 x) x x^\top \right\rangle_{\mathcal{D}} \right] && (5)\\
\dot{W}_2 &= \left\langle \left[y - W_2 \text{max}(0, W_1 x) \right] \text{max}(0, W_1 x)^\top \right\rangle_{\mathcal{D}} && (6)\\
&= \left\langle \left[y - W_2 \theta(W_1 x) W_1 x \right] \theta(W_1 x) x^\top W_1^\top \right\rangle_{\mathcal{D}} && (7)\\
&= \left[\Sigma^{yx} - W_2 W_1 \Sigma^x \right] W_1^\top && (8)
\end{aligned}$$

Therefore these dynamics look like those of a deep linear network but with input-output correlations $\Sigma^{yx} = \left\langle \theta(W_1 x) y x^\top \right\rangle_{\mathcal{D}}, \Sigma^x = \left\langle \theta(W_1 x) x x^\top \right\rangle_{\mathcal{D}}$.

2. Suppose the first layer weights are initialized to $W_1(0) = \begin{bmatrix} 0 & 0 & \alpha \end{bmatrix}$ with $\alpha > 0$. What are the effective dataset statistics?

   (2 marks)

   Solution

   The effective dataset contains those samples for which $W_1 x > 0$. We therefore have

$$\begin{aligned}
\Sigma^{yx} &= \frac{1}{4} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ \Delta & \Delta \end{bmatrix}^\top && (9)\\
&= \begin{bmatrix} 0 & 0 & \frac{\Delta}{2} \end{bmatrix} && (10)\\
\Sigma^x &= \frac{1}{4} \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ \Delta & \Delta \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ \Delta & \Delta \end{bmatrix}^\top && (11)\\
&= \frac{1}{2} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & \Delta^2 \end{bmatrix}. && (12)
\end{aligned}$$

3. If $\Delta > 0$ and the weights are initialized such that $W_2(0) > 0, W_1(0) = \begin{bmatrix} 0 & 0 & \alpha \end{bmatrix}$ with $\alpha > 0$, show that under the gradient flow dynamics the first layer weights remain in this form, that is, $W_1(t) = \begin{bmatrix} 0 & 0 & \alpha(t) \end{bmatrix}$ with $\alpha(t) > 0$ for all t.

   (8 marks)

Solution We know

$$\dot{W}_1 = W_2\left(\Sigma^{yx} - W_2 W_1 \Sigma^x\right) \tag{13}$$

$$= W_2\left(\begin{bmatrix}0 & 0 & \frac{\Delta}{2}\end{bmatrix} - W_2\begin{bmatrix}0 & 0 & \alpha\end{bmatrix}\frac{1}{2}\begin{bmatrix}1 & -1 & 0\\ -1 & 1 & 0\\ 0 & 0 & \Delta^2\end{bmatrix}\right). \tag{14}$$

$$= \begin{bmatrix}0 & 0 & \frac{1}{2}(W_2\Delta - W_2^2\alpha\Delta^2)\end{bmatrix}. \tag{15}$$

$$\dot{W}_2 = \left[\Sigma^{yx} - W_2 W_1 \Sigma^x\right] W_1^\top \tag{16}$$

$$= \begin{bmatrix}0 & 0 & \frac{1}{2}(\Delta - W_2\alpha\Delta^2)\end{bmatrix}\begin{bmatrix}0 & 0 & \alpha\end{bmatrix}^\top \tag{17}$$

$$= \frac{1}{2}(\Delta\alpha - W_2\alpha^2\Delta^2). \tag{18}$$

From this we see that W_1 will have only its third element nonzero. We must now show that W_1 cannot change sign. We have the two dimensional dynamics

$$\dot{\alpha} = \frac{1}{2}(W_2\Delta - W_2^2\alpha\Delta^2) \tag{19}$$

$$\dot{W}_2 = \frac{1}{2}(\Delta\alpha - W_2\alpha^2\Delta^2). \tag{20}$$

Consider the ray where $\alpha = 0, W_2 \geq 0$. The derivative along this ray is

$$\dot{\alpha} = \frac{1}{2}W_2\Delta \tag{21}$$

$$\dot{W}_2 = 0, \tag{22}$$

such that the dynamics flow across this ray into the positive quadrant, since $\frac{1}{2}W_2\Delta \geq 0$. Now consider the ray where $\alpha \geq 0, W_2 = 0$, yielding derivative

$$\dot{\alpha} = 0 \tag{23}$$

$$\dot{W}_2 = \frac{1}{2}\Delta\alpha, \tag{24}$$

such that the dynamics again flow into the positive quadrant, since $\frac{1}{2}\Delta\alpha > 0$. Since these two rays trace the boundary of the positive quadrant and trajectories never cross out of the boundaries, trajectories are confined to the positive quadrant and cannot change the sign of W_1.

4. Recall that the dynamics of a deep linear network depend on the singular value decomposition of the input-output correlations $\Sigma^{yx} = USV^\top$ where S is diagonal. When initialized with small random weights, each singular value mode is learned in time approximately order $O(1/s)$, ignoring logarithmic factors. Using this fact and the preceding sections, what is the approximate learning time for a single hidden neuron ReLU network initialized with $W_2(0) > 0$, $W_1(0) = \begin{bmatrix}0 & 0 & \alpha\end{bmatrix}$ and $\alpha > 0$?

   (5 marks)

   Solution Because we have shown that the ReLU network behaves like a linear network on a specific effective dataset, its learning time is the learning time of the linear network. We therefore need to find the singular value s.

$$\Sigma^{yx} = \begin{bmatrix}0 & 0 & \frac{\Delta}{2}\end{bmatrix} = \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}\frac{\Delta}{2}\end{bmatrix}\begin{bmatrix}0 & 0 & 1\end{bmatrix} \tag{25}$$

so the singular value is $s = \Delta/2$ and the approximate learning time is $t \approx 2/\Delta$.

5. If instead we have a neuron with weights proportional to the first training input sample,

$$W_1 = \alpha[-1 - 1 - \Delta] \tag{26}$$

for $\alpha > 0$, and $0 < \Delta < \sqrt{2}$, what is this neuron's effective dataset and approximate learning speed? (You do not need to show that the dynamics remain proportional to this input sample, though it turns out they do.)

(7.5 marks)

Solution

The effective dataset only contains this one input sample because the dot product with the other three examples is $-\Delta^2, -\Delta^2$ and $-2 + \Delta^2$ respectively. These are all less than zero under the assumptions.

The singular value of $\Sigma^{yx} = \frac{1}{4}\begin{bmatrix}1 & 1 & \Delta\end{bmatrix}$ is $s = \sqrt{\frac{1}{8} + \frac{\Delta^2}{16}}$.

You have calculated the speed at which a neuron exploiting linear separability and a neuron that 'memorizes' one example would learn, if these were the only neurons in a network. Now consider having many neurons. Intuitively, at initialization, some will exploit linear separability while others will be in the memorizing configuration. There will be an approximate race between these different neurons. We might expect the resulting dynamics to be dominated by whichever neuron type is faster.

6. For what range of Δ will the linear separability neurons learn faster? For what range of Δ will the 'memorizing' neurons learn faster? What is the critical Δ separating these regimes? According to this analysis, does a ReLU network trained with gradient descent exploit linear separability whenever it can?

(7.5 marks)

Solution The crossover point is

$$\Delta/2 = \sqrt{\frac{1}{8} + \frac{\Delta^2}{16}} \tag{27}$$

$$\Delta^2/4 = \frac{1}{8} + \frac{\Delta^2}{16} \tag{28}$$

$$4\Delta^2 = 2 + \Delta^2 \tag{29}$$

$$\Delta = \sqrt{2/3}. \tag{30}$$

By testing points, we see that for $0 \leq \Delta < \sqrt{2/3}$ the memorization solution dominates while for $\Delta > \sqrt{2/3}$ the linear separability solution dominates. Hence a ReLU network does not always exploit linear separability. The linear margin is 2Δ but for small enough margins, the race will be won by memorizing solutions.