M2 Biology **3/12/20**

TD8: Learning II

Elie Oriol

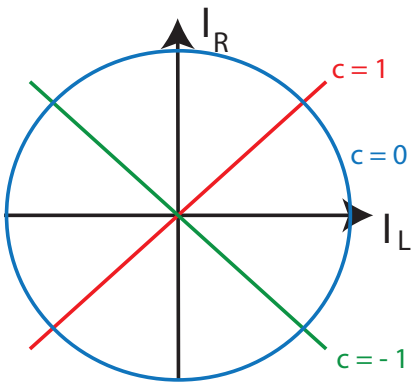
https://github.com/Elieoriol/2021_UlmM2_ThNeuro/tree/master/TD8

Following on our series of tutorials on learning, this one is devoted to unsupervised learning.

1 Inputs

We consider a neuron receiving two inputs, for example visual input from the left eye I_L and visual input from the right eye I_R. Each input is drawn from a random distribution of mean 0 and variance v. Moreover, the two inputs are correlated according to: $\langle I_L I_R \rangle = c$.

1. For $v = 1$ and $c = -1, 0, 1$, sketch a distribution where each input varies between -1 and 1. Explain why for visual input we should have $c \geq 0$.



For visual inputs, both eyes see the same visual world from slightly different viewpoints; what they see is therefore correlated, not anti-correlated.

2. Show that $-v \leq c \leq v$.

$$\langle (I_L - I_R)^2 \rangle = \langle I_L^2 \rangle - 2\langle I_L I_R \rangle + \langle I_R^2 \rangle = 2(v - c) > 0$$
$$\langle (I_L + I_R)^2 \rangle = \langle I_L^2 \rangle + 2\langle I_L I_R \rangle + \langle I_R^2 \rangle = 2(v + c) > 0$$

3. What are the correlation and anti-correlation axes $\vec{e}_1, \vec{e}_2$ of the distribution? Write them as a function of the basis vectors $\vec{e}_L, \vec{e}_R$. For any vector $\vec{X} = x_L \vec{e}_L + x_R \vec{e}_R$, what are the corresponding coordinates in the new basis $\vec{X} = x_1 \vec{e}_1 + x_2 \vec{e}_2$?

$$\vec{e}_1 = \frac{\vec{e}_L + \vec{e}_R}{\sqrt{2}} \quad \vec{e}_2 = \frac{\vec{e}_L - \vec{e}_R}{\sqrt{2}} \qquad \Rightarrow \qquad x_1 = \frac{x_L + x_R}{\sqrt{2}} \quad x_2 = \frac{x_L - x_R}{\sqrt{2}}$$

4. Calculate the correlations $\langle I_1^2 \rangle$, $\langle I_2^2 \rangle$, $\langle I_1 I_2 \rangle$.

   For $\vec{X} = I_1$ and I_2:

$$\begin{aligned} \langle I_1^2 \rangle &= \langle \frac{(I_L + I_R)^2}{2} \rangle = \frac{\langle I_L^2 \rangle + \langle I_R^2 \rangle + 2\langle I_L I_R \rangle}{2} = \frac{v + v + 2c}{2} = v + c \\ \langle I_1^2 \rangle &= \langle \frac{(I_L - I_R)^2}{2} \rangle = \frac{\langle I_L^2 \rangle + \langle I_R^2 \rangle - 2\langle I_L I_R \rangle}{2} = \frac{v + v - 2c}{2} = v - c \\ \langle I_1 I_2 \rangle &= \langle \frac{(I_L - I_R)(I_L + I_R)}{2} \rangle = \frac{\langle I_L^2 \rangle - \langle I_R^2 \rangle}{2} = \frac{v - v}{2} = 0 \end{aligned}$$

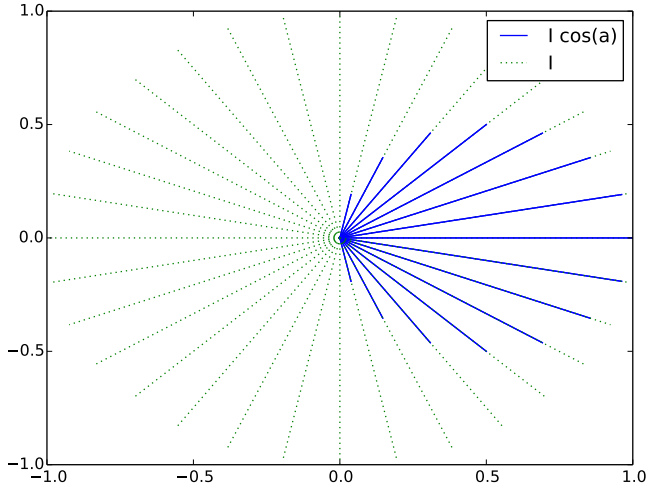
2 Hebbian learning algorithm

The activity of the neuron is given by $V = \vec{W} \cdot \vec{I}$. We consider a Hebbian learning rule in which, every time an input $\vec{I}(t)$ is presented, the neuron weights are updated according to:

$$\vec{W}(t+1) = \vec{W}(t) + \epsilon V(t) \vec{I}(t) \tag{1}$$

5. We denote α the angle between $\vec{I}(t)$ and $\vec{W}(t)$. Supposing that $\|\vec{I}\| = 1$, sketch the update as a function of α. How does $\|\vec{W}\|$ evolve?

$$\vec{W}(t+1) - \vec{W}(t) = \epsilon V(t) \vec{I}(t) = \epsilon \|\vec{W}\| \cos(\alpha) \vec{I}$$



Considering different angles α, one can see that $\|\vec{W}\|$ increases at each update.

To make more precise statements, we remove ϵ to simplify and study the mean dynamics:

$$\frac{\mathrm{d}\vec{W}}{\mathrm{d}t} = \langle V(t) \vec{I}(t) \rangle \tag{2}$$

where the average $\langle \cdot \rangle$ is taken over the distribution of the inputs $\vec{I}$.

6. For each of the axes $\vec{e}_1$, $\vec{e}_2$ of the input distribution, suppose that initially $\vec{W}$ is along one of this direction, then in which direction is $\frac{\mathrm{d}\vec{W}}{\mathrm{d}t}$? Along which of these two directions will $\frac{\mathrm{d}\vec{W}}{\mathrm{d}t}$ have the largest magnitude?

   If $\vec{W}$ is along one of the axes e_1, e_2, then through averaging $\frac{\mathrm{d}\vec{W}}{\mathrm{d}t}$ will be parallel to $\vec{W}$. Let's take for instance $\vec{W} = w\vec{e_1}$:

$$\frac{\mathrm{d}\vec{W}}{\mathrm{d}t} = w\langle(\vec{e_1}\cdot\vec{I})\vec{I}\rangle = w\left[\langle I_1^2\rangle\vec{e_1} + \langle I_1 I_2\rangle\vec{e_2}\right] = (v+c)\vec{W}$$

   The axes $\vec{e}_1$, $\vec{e}_2$ will therefore be eigenvectors of the dynamics. $\vec{e}_1$ will be associated to the largest eigenvalue $v+c$, $\vec{e}_2$ with $v-c$.

7. Obtain a linear differential equation on $\vec{W}$. What are the eigenvectors and associated eigenvalues? Comment on the dynamics.

   In the $\vec{e}_L, \vec{e}_R$ basis:

$$\begin{aligned}\frac{\mathrm{d}\vec{W}}{\mathrm{d}t} &= \langle V(t)\vec{I}(t)\rangle = \langle\begin{pmatrix} W_L I_L^2 + W_R I_L I_R \\ W_R I_R^2 + W_L I_L I_R \end{pmatrix}\rangle \\ &= \begin{pmatrix} W_L v + W_R c \\ W_R v + W_L c \end{pmatrix} = \begin{pmatrix} v & c \\ c & v \end{pmatrix}\vec{W} = A\vec{W}\end{aligned}$$

   as the mean of a vector is the vector of the means. For those not familiar with it, this differential equation is a matricial one. $\vec{W}$ is the vector $\begin{pmatrix} W_L \\ W_R \end{pmatrix}$ in the $\vec{e}_L, \vec{e}_R$ basis. The time derivative of this vector is equal to the product of the matrix A with $\vec{W}$. Projecting on each axis, this is totally equivalent to:

$$\begin{aligned}\frac{\mathrm{d}W_L}{\mathrm{d}t} &= W_L v + W_R c \\ \frac{\mathrm{d}W_R}{\mathrm{d}t} &= W_R v + W_L c\end{aligned}$$

   It is useful to diagonalize A to get the linear evolution of $\vec{W}$ on each of the orthogonal axes defined by the eigenvectors of A (which we expect to be e_1, e_2 from the previous question). The eigenvalues are given by finding the roots of the characteristic polynomial:

$$\begin{aligned}\det[A - \lambda I_2] &= 0 \\ (v-\lambda)^2 - c^2 &= 0 \\ v - \lambda &= \pm c\end{aligned}$$

   The two solutions are $\lambda_1 = v + c$ and $\lambda_2 = v - c$.

   The normalised eigenvector $\vec{e_1}$ associated to the first eigenvalue satisfies:

$$A\vec{e_1} = \vec{e_1}$$

   Writing $\vec{e_1} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ and normalizing at the end:

$$\begin{pmatrix} v & c \\ c & v \end{pmatrix}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} vx_1 + cy_1 \\ cx_1 + vy_1 \end{pmatrix} = \begin{pmatrix} (v+c)x_1 \\ (v+c)y_1 \end{pmatrix}$$

$$\Rightarrow x_1 = y_1 \Rightarrow \vec{e_1} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

The same development for $\vec{e_2}$ leads to:

$$\begin{pmatrix} v & c \\ c & v \end{pmatrix}\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} vx_2 + cy_2 \\ cx_2 + vy_2 \end{pmatrix} = \begin{pmatrix} (v-c)x_2 \\ (v-c)y_2 \end{pmatrix}$$
$$\Rightarrow x_2 = -y_2 \Rightarrow \vec{e_2} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

Using the previous question, we could have considered the dynamics in the $\vec{e}_1, \vec{e}_2$ basis from the start, in which case, writing now $\vec{W} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$:

$$\frac{\mathrm{d}\vec{W}}{\mathrm{d}t} = \langle V(t)\vec{I}(t)\rangle = \langle \begin{pmatrix} W_1 I_1^2 + W_2 I_1 I_2 \\ W_2 I_2^2 + W_1 I_1 I_2 \end{pmatrix}\rangle$$
$$= \begin{pmatrix} W_1(v+c) + 0 \\ 0 + W_2(v-c) \end{pmatrix} = \begin{pmatrix} v+c & 0 \\ 0 & v-c \end{pmatrix}\vec{W} = D\vec{W}$$

We see that now the system is already diagonalized; considering the evolution directly in the eigenvectors basis, the matrix of evolution D is diagonal. The previous equation is as such equivalent to:

$$\frac{\mathrm{d}W_1}{\mathrm{d}t} = (v+c)W_1$$
$$\frac{\mathrm{d}W_2}{\mathrm{d}t} = (v+c)W_2$$

8. We add a "homeostatic" term to the dynamics so as to prevent the weights from growing exponentially:

$$\frac{\mathrm{d}\vec{W}}{\mathrm{d}t} = \langle V(t)\vec{I}(t)\rangle - \langle V(t)^2\rangle\vec{W}(t) \tag{3}$$

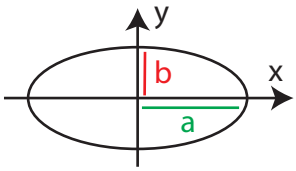
Can we obtain a **linear** differential equation on $\vec{W}$? Obtain a differential equation on $\vec{W}$ in the basis $(\vec{e_1}, \vec{e_2})$.

The second term includes components of $\vec{W}$ to the power three, we therefore cannot obtain a linear differential equation.

$$\langle V(t)^2\rangle = \langle (W_1 I_1 + W_2 I_2)^2\rangle = W_1^2\langle I_1^2\rangle + W_2^2\langle I_2^2\rangle + W_1 W_2\langle I_1 I_2\rangle = W_1^2(v+c) + W_2^2(v-c)$$
$$\frac{\mathrm{d}W_1}{\mathrm{d}t} = (v+c)W_1 - W_1\langle V(t)^2\rangle = W_1(v + c - W_1^2(v+c) - W_2^2(v-c))$$
$$\frac{\mathrm{d}W_2}{\mathrm{d}t} = (v-c)W_2 - W_2\langle V(t)^2\rangle = W_2(v - c - W_2^2(v-c) - W_1^2(v+c))$$

9. Draw the nullclines. For this you will need the equation of an ellipse, given by:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{4}$$



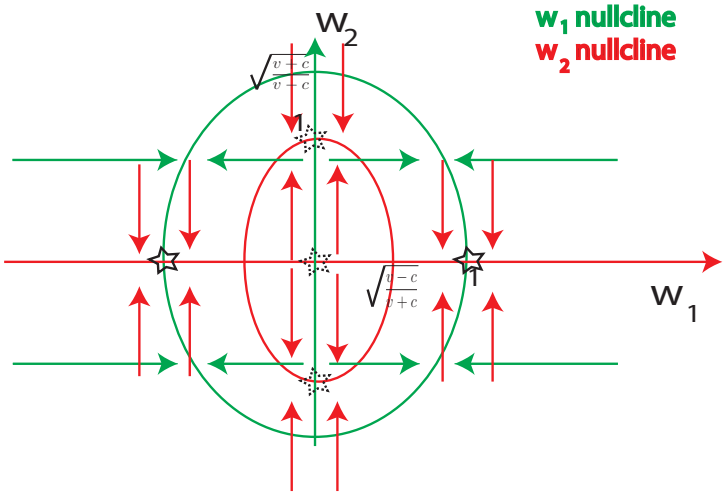
What are the equilibrium points for $\vec{W}$? Are they stable?

We rewrite the equations to make the ellipses appear:

$$\frac{\mathrm{d}W_1}{\mathrm{d}t} = -(v+c)\,W_1\left(W_1^2 + \frac{v-c}{v+c}W_2^2 - 1\right)$$
$$\frac{\mathrm{d}W_2}{\mathrm{d}t} = -(v-c)\,W_2\left(\frac{v+c}{v-c}W_1^2 + W_2^2 - 1\right)$$

The nullclines correspond to the points where $\frac{\mathrm{d}W_1}{\mathrm{d}t}$ *and* $\frac{\mathrm{d}W_2}{\mathrm{d}t}$ *cancel.*

Therefore, the W_1 *nullcline contains the line* $W_1 = 0$ *and the ellipse* $W_1^2 + \frac{v-c}{v+c}W_2^2 = 1$*, i.e. the ellipse with* $a_1 = 1, b_1 = \sqrt{\frac{v+c}{v-c}}$*. The* W_2 *nullcline contains the line* $W_2 = 0$ *and the ellipse* $\frac{v+c}{v-c}W_1^2 + W_2^2 = 1$*, i.e. the ellipse with* $a_2 = \sqrt{\frac{v-c}{v+c}}, b_1 = 1$*.*



The previous figure might seem complicated but can be simply understood: the arrows indicate the direction of evolution of the system in a given part of the (W_1, W_2) *space. To understand this, one should take a point anywhere in the space, and depending on its position relative to the different nullclines, see if at this point the derivatives of* W_1 *and* W_2*, obtained at the previous question are positive or negative. A positive derivative for* W_1 *means a right arrow and for* W_2 *an up arrow, a negative one for* W_1 *means a left arrow and for* W_2 *a down arrow.*

The only stable intersection points are $W_2 = 0, W_1 = \pm 1$*. Since initially the weights are positive,* $W_1(t=0) > 0$*, the system converges to* $W_1 = 1, W_2 = 0$*. This means we only keep the projection onto the principal component, that is the eigenvector associated to the largest eigenvalue. The projection on the second eigenvector is discarded, made equal to* 0*.*

10. We would now like to study competitive Hebbian learning. We add a term to the dynamics so as to introduce competition between the left and right inputs. In the $(\vec{e_L}, \vec{e_R})$, the dynamics are now given by:

$$\frac{\mathrm{d}\vec{W}}{\mathrm{d}t} = \langle V(t)\vec{I}(t)\rangle - \langle V(t)\begin{pmatrix} \frac{I_L+I_R}{2} \\ \frac{I_L+I_R}{2} \end{pmatrix}\rangle \tag{5}$$

Obtain a linear differential equation on $\vec{W}$ in the basis $(\vec{e_1}, \vec{e_2})$. Comment on the dynamics.

The learning rule is now:

$$\begin{aligned}\frac{\mathrm{d}\vec{W}}{\mathrm{d}t} &= \langle V(t)\vec{I}(t)\rangle - \langle V(t)I_1\vec{e}_1\rangle = \langle V(t)(I_1\vec{e}_1 + I_2\vec{e}_2 - I_1\vec{e}_1)\rangle \\ &= \langle V(t)I_2\vec{e}_2\rangle = \langle (W_1I_1 + W_2I_2)I_2\rangle\vec{e}_2 = W_2(v-c)\vec{e}_2 \\ &= \begin{pmatrix} 0 & 0 \\ 0 & v-c \end{pmatrix}\vec{W}\end{aligned}$$

The W_1 component doesn't change and the W_2 component grows exponentially, towards $\pm\infty$ depending on the sign of W_2.

11. We enforce the weights to remain positive. Show that, depending on the initial conditions, only one of the two weights will be non-zero.

    From the previous question we know $W_L + W_R$ is constant, and $W_L - W_R$ is growing exponentially. Enforcing both to be positive, their sum is fixed and their difference goes exponentially to $+\infty$ if initially $W_L > W_R$ or to $-\infty$ in the other case. This means the highest initially wins, the other becoming null.