M2 Cogmaster 2025 | October 23$^{rd}$
TA: Yoan Fardella (yoan.fardella@ens.psl.eu)

ENS | PSL

**THEORETICAL NEUROSCIENCE**

**TD7: UNSUPERVISED LEARNING**

All TD materials will be made available at `https://github.com/yfardella/Th_Neuro_TD_2025`.

Several learning paradigms exist: mainly supervised, unsupervised and reinforcement learning. In these series of tutorials, we will develop standard models in each of these three paradigms. This second tutorial of these series presents unsupervised learning through the paradigm of Hebbian learning, which is a biologically plausible mechanism in neurons.

## 1 Binocular Neuron

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 $\mathrm{Cov}(I_L, I_R) = c$.

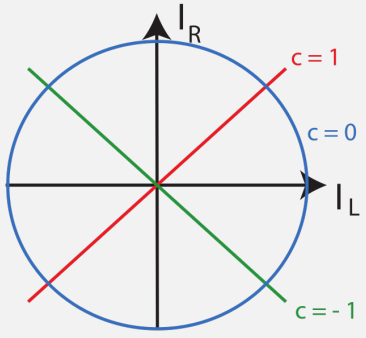
*Reminder*: for two random variables $X$ and $Y$,

$$\mathbb{C}\mathrm{ov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \text{ and } \mathbb{C}\mathrm{orr}(X,Y) = \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$

1. For $v = 1$ and $c \in \{-1, 0, 1\}$, sketch a distribution in the plane $(I_L, I_R)$ where each input varies between $-1$ and 1.

Different inputs correspond to points in the plane $(I_L, I_R)$. The correlation sets the directions along which inputs align.

- If $c = 0$ then inputs are not correlated. Thus, they can span the full unit disk.
- If $c = 1$ then inputs are positively correlated, such that high values of $I_L$ are associated with high values of $I_R$. Thus, inputs lie in an ellipse along the identity line.
- If $c = -1$ then inputs are negatively correlated, such that positive values of $I_L$ are associated to negative values of $I_R$. Thus, inputs lie in an ellipse along the line $y = -x$.



2. Justify why, for visual inputs, the correlation between left and right inputs should be modelled as $c \geq 0$.

> Both eyes receive light from the same visual scene, with slightly different viewpoints. Thus, left and right inputs are positively correlated, which means that $c \geq 0$.

3. Justify that $\mathbb{C}\text{ov}(I_L, I_R) = \mathbb{E}(I_L I_R)$, $\mathbb{V}(I_L) = \mathbb{E}(I_L^2)$ and $\mathbb{V}(I_R) = \mathbb{E}(I_R^2)$.

> By definition of covariance and variance, and using the fact that the inputs are drawn from a distribution of mean zero we have that:
>
> $$\mathbb{C}\text{ov}(I_L, I_R) = \mathbb{E}[(I_L - \mathbb{E}(I_L))(I_R - \mathbb{E}(I_R))] = \mathbb{E}(I_L I_R) - \mathbb{E}(I_L)\mathbb{E}(I_R) = \mathbb{E}(I_L I_R),$$
>
> $$\mathbb{V}(I_L) = \mathbb{E}(I_L^2) - \mathbb{E}(I_L)^2 = \mathbb{E}(I_L^2) \text{ and } \mathbb{V}(I_R) = \mathbb{E}(I_R^2) - \mathbb{E}(I_R)^2 = \mathbb{E}(I_R^2).$$

4. Prove that $-v \leq \mathbb{C}\text{ov}(I_L, I_R) = c \leq v$ and deduce that $-1 \leq \mathbb{C}\text{orr}(I_L, I_R) \leq 1$.
*Hint:* one can use the previous question or the Cauchy-Schwartz inequality.

> Method 1: using the previous question, the relation between the variance and the covariance can be obtained by developing the square of the sum and the difference of the random variables $I_R$ and $I_L$. Since cubes are always positive, we have that:
>
> $$\mathbb{E}((I_L - I_R)^2) = \mathbb{E}(I_L^2) - 2\mathbb{E}(I_L I_R) + \mathbb{E}(I_R^2) = 2(v - c) \geq 0 \Rightarrow v \geq c,$$
>
> $$\mathbb{E}((I_L + I_R)^2) = \mathbb{E}(I_L^2) + 2\mathbb{E}(I_L I_R) + \mathbb{E}(I_R^2) = 2(v + c) \geq 0 \Rightarrow -v \leq c,$$
>
> and therefore:
>
> $$-v \leq c \leq v.$$
>
> Method 2: using the Cauchy-Schwartz inequality $\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$ applied to $X = I_L - \mathbb{E}(I_L)$ and $Y = I_R - \mathbb{E}(I_R)$, we have that:
>
> $$\underbrace{\mathbb{E}[(I_L - \mathbb{E}(I_L))(I_R - \mathbb{E}(I_R))]^2}_{:= \mathbb{C}\text{ov}(I_L, I_R)^2 = c^2} \leq \underbrace{\mathbb{E}[(I_L - \mathbb{E}(I_L))^2]}_{:= \mathbb{V}(I_L) = v} \underbrace{\mathbb{E}[(I_R - \mathbb{E}(I_R))^2]}_{:= \mathbb{V}(I_R) = v},$$
>
> and therefore:
>
> $$-v \leq c \leq v.$$
>
> Furthermore:
>
> $$\mathbb{C}\text{orr}(I_L, I_R) = \frac{c}{v} \Rightarrow -1 \leq \mathbb{C}\text{orr}(I_L, I_R) \leq 1.$$

5. Let $(\vec{e}_1, \vec{e}_2)$ be two basis vectors (with unit norm) aligned with the axes of perfect correlation and perfect anti-correlation. Express those basis vectors as combinations of the canonical basis vectors $(\vec{e}_L, \vec{e}_R)$.

> The axis reflecting perfect correlation is along the vector $(1, 1)^T$ with unit norm:
>
> $$\vec{e}_1 = \frac{1}{\sqrt{2}} \left( \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) = \frac{1}{\sqrt{2}} (\vec{e}_L + \vec{e}_R).$$

The axis reflecting perfect anti-correlation is along the vector $(1,-1)^T$ with unit norm:

$$\vec{e}_2 = \frac{1}{\sqrt{2}}\left(\begin{bmatrix}1\\0\end{bmatrix} - \begin{bmatrix}0\\1\end{bmatrix}\right) = \frac{1}{\sqrt{2}}(\vec{e}_L - \vec{e}_R).$$

Each input can be decomposed in the canonical basis $\vec{I} = (I_L, I_R)$ or equivalently in the new basis $\vec{I} = (I_1, I_2)$. The coefficients $I_L, I_R, I_1$ and $I_2$ correspond to random variables related between each other.

6. For any vector $\vec{I} = I_L\vec{e}_L + I_R\vec{e}_R$, express its coordinates in the new basis $\vec{I} = I_1\vec{e}_1 + I_2\vec{e}_2$.

From the previous question, we have that:

$$\vec{e}_1 = \frac{\vec{e}_L + \vec{e}_R}{\sqrt{2}} \text{ and } \vec{e}_2 = \frac{\vec{e}_L - \vec{e}_R}{\sqrt{2}} \;\Rightarrow\; \vec{e}_L = \frac{\vec{e}_1 + \vec{e}_2}{\sqrt{2}} \text{ and } \vec{e}_R = \frac{\vec{e}_1 - \vec{e}_2}{\sqrt{2}}.$$

Therefore:

$$\vec{I} = I_L\vec{e}_L + I_R\vec{e}_R = I_L\frac{\vec{e}_1 + \vec{e}_2}{\sqrt{2}} + I_R\frac{\vec{e}_1 - \vec{e}_2}{\sqrt{2}} = \underbrace{\frac{I_L + I_R}{\sqrt{2}}}_{:=I_1}\vec{e}_1 + \underbrace{\frac{I_L - I_R}{\sqrt{2}}}_{:=I_2}\vec{e}_2.$$

7. Compute $\mathbb{E}(I_1^2), \mathbb{E}(I_2^2)$ and $\mathbb{E}(I_1 I_2)$.

We have that:

$$\mathbb{E}(I_1^2) = \mathbb{E}\left(\left(\frac{I_L + I_R}{\sqrt{2}}\right)^2\right) = \frac{\mathbb{E}(I_L^2) + 2\mathbb{E}(I_L I_R) + \mathbb{E}(I_R^2)}{2} = \frac{v + 2c + v}{2} = v + c,$$

$$\mathbb{E}(I_2^2) = \mathbb{E}\left(\left(\frac{I_L - I_R}{\sqrt{2}}\right)^2\right) = \frac{\mathbb{E}(I_L^2) - 2\mathbb{E}(I_L I_R) + \mathbb{E}(I_R^2)}{2} = \frac{v - 2c + v}{2} = v - c,$$

$$\mathbb{E}(I_1 I_2) = \mathbb{E}\left(\frac{(I_L + I_R)(I_L - I_R)}{2}\right) = \frac{\mathbb{E}(I_L^2) - \mathbb{E}(I_R^2)}{2} = \frac{v - v}{2} = 0.$$

## 2 Hebbian Learning

In the model, the activity of the binocular neuron is a linear combination of the random inputs it receives at any time:

$$V(t) = \vec{W}(t) \cdot \vec{I}(t). \tag{1}$$

The weights $W$ represent the synaptic strengths between the sensory/retina neurons and the binocular neuron.

In unsupervised learning, the synaptic weights evolve depending only on the neuron's activity itself. The *learning rule* specifies the update of the weights' vector $\vec{W}$ every time an input $\vec{I}(t)$ is presented, under the form:

$$\vec{W}(t+1) = \vec{W}(t) + f(V(t), \vec{I}(t)). \tag{2}$$

Several learning rules exist, implementing different choices for the update function $f$. Most of them are variants of the standard Hebbian learning rule presented below.

## 2.1 Standard Hebbian Learning

According to the Hebbian learning rule, every time an input $\vec{I}(t)$ is presented, the neuron weights are updated according to:

$$\vec{W}(t+1) = \vec{W}(t) + \epsilon V(t)\vec{I}(t). \tag{3}$$

We study the mean dynamics:

$$\frac{d\vec{W}}{dt} = \epsilon \langle V(t)\vec{I}(t) \rangle,$$

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

8. Let $\alpha$ denote the angle between $\vec{I}(t)$ and $\vec{W}(t)$. Assuming $\|\vec{I}\| = 1$, sketch the update of the vector $\vec{W}$ in the plane $(w_1, w_2)$ for different values of $\alpha$. Comment on the evolution of $\|\vec{W}\|$.

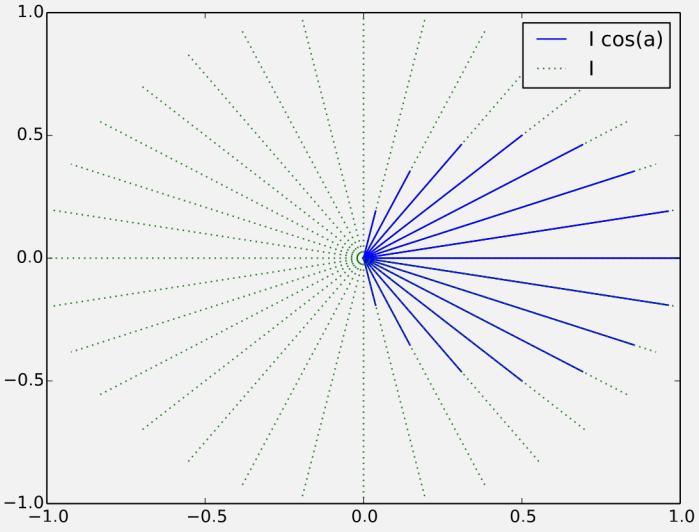
With the Hebbian learning rule, $\vec{W}(t+1) - \vec{W}(t) = \epsilon V(t)\vec{I}(t)$ such that the update vector is aligned in the direction of the input $\vec{I}(t)$ with magnitude $\epsilon V(t)$ (under the assumption that $\|\vec{I}\| = 1$.)

Moreover, the activity of the neuron is exactly the scalar product $V(t) = \vec{W}(t) \cdot \vec{I}(t)$. This is equivalent (*bonus: prove that the algebraic and geometric definitions of a scalar product are equivalent*) to $V(t) = \|\vec{W}\|\|\vec{I}\| \cos(\vec{W}, \vec{I}) = \|\vec{W}\| \cos(\alpha)$.

Therefore, the update vector can be expressed as:

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

Furthermore, whatever the angle $\alpha$, one can see that the norm of the weight vector $\|\vec{W}\|$ increases after each update.



In what follows, the learning rate is set to $\epsilon = 1$ to simplify the next questions.

9. In the case in which $\vec{W}$ is initially along the direction of one main axis of the input distribution, $\vec{e}_1$ or $\vec{e}_2$, determine the corresponding direction of the update $\frac{d\vec{W}}{dt}$. Along which of these two directions would the update vector have the largest magnitude?

> If $\vec{W}$ is along one of the axes $\vec{e}_1, \vec{e}_2$, then through averaging, the update vector $\frac{d\vec{W}}{dt}$ will be parallel to $\vec{W}$. For instance, with $\vec{W} = w\vec{e}_1$:
>
> $$\begin{aligned}\frac{d\vec{W}}{dt} &= \langle V(t)\vec{I}(t)\rangle = \langle \vec{W}(t)\cdot\vec{I}(t)\times\vec{I}(t)\rangle = \langle w\vec{e}_1\cdot\vec{I}(t)\times\vec{I}(t)\rangle = \langle wI_1\times(I_1\vec{e}_1+I_2\vec{e}_2)\rangle \\ &= w\left(\langle I_1^2\rangle\vec{e}_1 + \langle I_1I_2\rangle\vec{e}_2\right) = w\left((v+c)\times\vec{e}_1 + 0\times\vec{e}_2\right)) = w(v+c)\vec{e}_1 = (v+c)\vec{W}.\end{aligned}$$
>
> Therefore, the axes $\vec{e}_1, \vec{e}_2$ are the eigenvectors of the dynamics: $\vec{e}_1$ is associated to the largest eigenvalue $v + c$ and $\vec{e}_2$ is associated to the lowest eigenvalue $v - c$.

10. Obtain a linear differential equation for the evolution of the weight vector $\vec{W}$. Determine the eigenvectors and associated eigenvalues of the dynamics.

> Let $\vec{W} = (w_1, w_2)^T$. The dynamics can be expressed directly in the eigenvector basis $(\vec{e}_1, \vec{e}_2)$ :
>
> $$\frac{d\vec{W}}{dt} = \langle V(t)\vec{I}(t)\rangle = \left\langle \begin{bmatrix} w_1I_1^2 + w_2I_1I_2 \\ w_1I_1I_2 + w_2I_2^2 \end{bmatrix}\right\rangle = \begin{bmatrix} w_1(v+c)+0 \\ 0 + w_2(v-c)\end{bmatrix} = \begin{pmatrix} v+c & 0 \\ 0 & v-c\end{pmatrix}\vec{W}.$$
>
> The system is already diagonalised, such that the evolution of the weights are gouverned by:
>
> $$\frac{dw_1}{dt} = (v+c)w_1 \quad \text{and} \quad \frac{dw_2}{dt} = (v-c)w_2.$$

## 2.2 Improvements of Hebbian Learning

In order to prevent the weights from growing exponentially, it is possible to add a "homeostatic" term to the dynamics, such that:

$$\frac{d\vec{W}}{dt} = \langle V(t)\vec{I}(t)\rangle - \langle V(t)^2\rangle\vec{W}(t).$$

11. Is it possible to obtain a linear differential equation for the evolution of $\vec{W}$? Obtain a differential equation on the components of $\vec{W}$ in the basis $(\vec{e}_1, \vec{e}_2)$.

> It is not possible to obtain a linear differential equation for the evolution of $\vec{W}$ because the second term includes components of $\vec{W}$ to the power three:
>
> $$\begin{aligned}\langle V(t)^2\rangle &= \langle(\vec{W}\cdot\vec{I})^2\rangle \\ &= \langle(w_1I_1 + w_2I_2)^2\rangle \\ &= w_1^2\langle I_1^2\rangle + w_2^2\langle I_2^2\rangle + 2w_1w_2\langle I_1I_2\rangle\end{aligned}$$

$$= w_1^2(v+c) + w_2^2(v-c) + 0,$$

and

$$\langle V(t)^2 \rangle \vec{W} = (w_1^2(v+c) + w_2^2(v-c)) \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} w_1^3(v+c) + w_1 w_2^2(v-c) \\ w_1^2 w_2(v+c) + w_2^3(v-c) \end{bmatrix}.$$

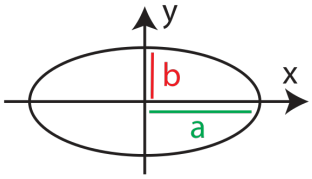
The evolution of the components of $\vec{W}$ obeys the following differential equations (terms in red correspond to the standard hebbian rule and those in blue to the homeostatic term):

$$\begin{cases} \frac{dw_1}{dt} = (v+c)w_1 - (w_1^3(v+c) + w_1 w_2^2(v-c)), \\ \frac{dw_2}{dt} = (v-c)w_2 - (w_1^2 w_2(v+c) + w_2^3(v-c)), \end{cases}$$

$$\Leftrightarrow \begin{cases} \frac{dw_1}{dt} = w_1(v + c - w_1^2(v+c) - w_2^2(v-c)), \\ \frac{dw_2}{dt} = w_2(v - c - w_2^2(v-c) - w_1^2(v+c)). \end{cases}$$

*Reminder:* The equation of an ellipse is given by:

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




12. Draw the nullclines in the space $(I_L, I_R)$. Identify the equilibrium points for $\vec{W}$, and determine their stability.

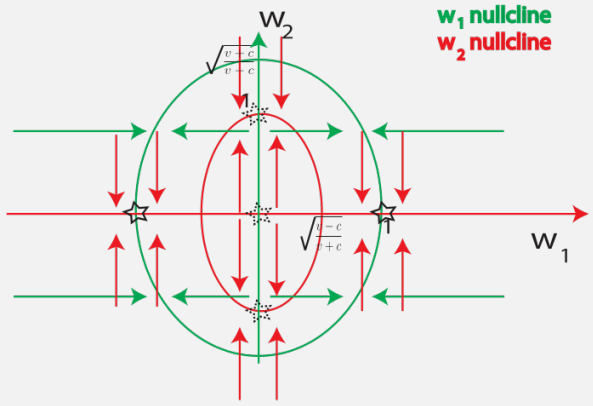
The differential equations can be rewritten under the form of the equation of ellipses:

$$\begin{cases} \frac{dw_1}{dt} = -(v+c)w_1 \left(w_1^2 + \frac{v-c}{v+c} w_2^2 - 1\right), \\ \frac{dw_2}{dt} = -(v-c)w_2 \left(\frac{v+c}{v-c} w_1^2 + w_2^2 - 1\right). \end{cases}$$

The nullclines correspond to the points where $\frac{dw_1}{dt}$ and $\frac{dw_2}{dt}$ cancel respectively:

- The $w_1$-nullcline contains:
  - the line $w_1 = 0$,
  - the ellipse $w_1^2 + \frac{v-c}{v+c} w_2^2 = 1$, i.e. the ellipse with $a_1 = 1, b_1 = \sqrt{\frac{v+c}{v-c}} > 1$.
- The $w_2$-nullcline contains:
  - the line $w_2 = 0$,
  - the ellipse $\frac{v+c}{v-c} w_1^2 + w_2^2 = 1$, i.e. the ellipse with $a_2 = \sqrt{\frac{v-c}{v+c}} < 1, b_2 = 1$.

In the figure below, the arrows indicate the direction of evolution of the system in any part of the $(w_1, w_2)$ space, which can be determined at any point depending on the sign of the derivatives of $w_1$ and $w_2$. A positive derivative entails a right arrow for $w_1$ and an up arrow for $w_2$ , and conversely.




The equilibria are obtained at the intersections of the $w_1-$ and $w_2-$nullclines. The only stable intersection points are $w_2 = 0, w_1 = \pm 1$.

13. Comment on the outcome of the homeostatic learning rule.

In both cases, at the equilibrium, the component $w_2$ is null. For instance, if initially the weights are positive, $w_1(t = 0) > 0$, then the system converges to $w_1 = 1, w_2 = 0$. This means the homeostatic learning rule keeps the projection onto the principal component, that is the eigenvector associated to the largest eigenvalue. The projection on the second eigenvector is discarded.

### 2.3 Competitive Hebbian Learning

Competitive Hebbian learning consists in adding a term to the dynamics so as to introduce competition between the left and right inputs. In the basis $(\vec{e}_1, \vec{e}_2)$, the dynamics are now given by:

$$\frac{d\vec{W}}{dt} = \langle V(t)\vec{I}(t)\rangle - \left\langle V(t) \begin{bmatrix} \frac{I_L+I_R}{2} \\ \frac{I_L+I_R}{2} \end{bmatrix} \right\rangle .$$

14. Obtain a linear differential equation on $\vec{W}$ in the basis $(\vec{e}_1, \vec{e}_2)$. Comment on the dynamics.

The second term of the learning rule can be expressed in the basis $(\vec{e}_1, \vec{e}_2)$ through the following relation:

$$\begin{bmatrix} \frac{I_L+I_R}{2} \\ \frac{I_L+I_R}{2} \end{bmatrix} = \frac{I_L + I_R}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{\sqrt{2}I_1}{2}\sqrt{2}\vec{e}_1 = I_1\vec{e}_1.$$

Therefore:

$$\left\langle V(t) \begin{bmatrix} \frac{I_L+I_R}{2} \\ \frac{I_L+I_R}{2} \end{bmatrix} \right\rangle = \langle (w_1 I_1 + w_2 I_2) I_1 \vec{e}_1 \rangle = (w_1 \langle I_1^2 \rangle + w_2 \langle I_2 I_1 \rangle)\vec{e}_1 = w_1(v + c)\vec{e}_1.$$

The evolution of the components of $\vec{W}$ obey the following differential equations (terms in red correspond to the standard hebbian rule and those in blue to the competitive term):

$$\begin{cases} \frac{dw_1}{dt} = (v+c)w_1 - (v+c)w_1, \\ \frac{dw_2}{dt} = (v-c)w_2. \end{cases} \iff \begin{cases} \frac{dw_1}{dt} = 0, \\ \frac{dw_2}{dt} = (v-c)w_2. \end{cases}$$

$$\iff \frac{d\vec{W}}{dt} = \begin{pmatrix} 0 & 0 \\ 0 & v-c \end{pmatrix} \vec{W}.$$

The component $w_1$ does not change and the component $w_2$ grows exponentially towards $\pm\infty$ depending on the sign of $w_2$.

15. If the weights are forced to remain positive, comment on the outcome of the competitive hebbian learning rule.

Returning in the initial basis, $w_L + w_R$ 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 weight with the highest initial value 'wins' the competition, while the other becomes null.