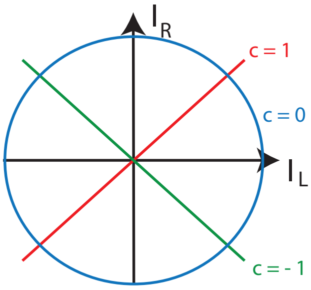
TD8 – Learning II Unsupervised Learning

1 Modeling inputs

① *Distributions of inputs*
Different inputs correspond to points in the plane (I_L, I_R). The correlation sets the direction 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 to 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 *Correlation for visual inputs*
Both eyes receive light from the same visual scene, with slightly different viewpoints. Thus, left and right inputs are correlated, which implies a positive correlation $c \geq 0$.

③ *Variances and Covariance*
According to Koenig-Huygens formula, for two random variables X, Y :

$$\mathbb{C}\text{ov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \qquad \mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

The means of the variables I_L and I_R are null by assumption, such that :

$$\mathbb{C}\text{ov}(I_L, I_R) = \mathbb{E}(I_L I_R) \qquad \mathbb{V}(I_L) = \mathbb{E}(I_L^2) \qquad \mathbb{V}(I_R) = \mathbb{E}(I_R^2)$$

④ *Ordering variances and covariances*
The relation between the variance and the covariance can be obtained by developing the square of the sum and the difference of the random variable I_R and I_L. As cubes are positive, so are the results :

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

Therefore :

$$-v \leq c \leq v$$

⑤ *Axes of maximal correlation and anti-correlation*

- The axis reflecting perfect correlation is along the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and has a unit norm : $\vec{e}_1 = \dfrac{\vec{e}_L + \vec{e}_R}{\sqrt{2}}$.
- The axis reflecting perfect anti-correlation is along the vector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and has a unit norm : $\vec{e}_2 = \dfrac{\vec{e}_L - \vec{e}_R}{\sqrt{2}}$.
- Equivalently, the vectors $\vec{e}_L$, $\vec{e}_R$ can be expressed in the basis $\vec{e}_1$, $\vec{e}_2$:

$$\begin{cases}\sqrt{2}\vec{e}_1 = \vec{e}_L + \vec{e}_R \\ \sqrt{2}\vec{e}_2 = \vec{e}_L - \vec{e}_R\end{cases} \implies \begin{cases}2\vec{e}_L = \sqrt{2}(\vec{e}_1 + \vec{e}_2) \\ 2\vec{e}_R = \sqrt{2}(\vec{e}_1 - \vec{e}_2)\end{cases} \implies \begin{cases}\vec{e}_L = \dfrac{\vec{e}_1 + \vec{e}_2}{\sqrt{2}} \\ \vec{e}_R = \dfrac{\vec{e}_1 - \vec{e}_2}{\sqrt{2}}\end{cases}$$

- For any vector $\vec{I} = I_L\vec{e}_L + I_R\vec{e}_R$, the corresponding coordinates in the new basis $\vec{I} = I_1\vec{e}_1 + I_2\vec{e}_2$ are :

$$I_1 = \frac{I_L + I_R}{\sqrt{2}} \qquad I_2 = \frac{I_L - I_R}{\sqrt{2}}$$

Those coordinates can be obtained by several methods.

Method ① Replacing the expressions of the vectors $\vec{e}_L$, $\vec{e}_R$ by $\vec{e}_1$, $\vec{e}_2$ and identifying.

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

Method ② Projecting the vector $\vec{I} = I_L\vec{e}_L + I_R\vec{e}_R$ onto the basis vectors $\vec{e}_1$ and $\vec{e}_2$ through the scalar product (since those basis vectors have unit norm and are orthogonal).

$$I_1 = \|\mathrm{Proj}_{\vec{e}_1}(\vec{I})\| = (I_L\vec{e}_L + I_R\vec{e}_R)\cdot\vec{e}_1 = I_L\vec{e}_L\cdot\vec{e}_1 + I_R\vec{e}_R\cdot\vec{e}_1 = I_L \times \frac{1}{\sqrt{2}} + I_R \times \frac{1}{\sqrt{2}} = \frac{I_L + I_R}{\sqrt{2}}$$
$$I_2 = \|\mathrm{Proj}_{\vec{e}_2}(\vec{I})\| = (I_L\vec{e}_L + I_R\vec{e}_R)\cdot\vec{e}_2 = I_L\vec{e}_L\cdot\vec{e}_2 + I_R\vec{e}_R\cdot\vec{e}_2 = I_L \times \frac{1}{\sqrt{2}} + I_R \times \left(-\frac{1}{\sqrt{2}}\right) = \frac{I_L - I_R}{\sqrt{2}}$$

⑥ *Correlations in the new basis*

The coefficients I_1 and I_2 can be expressed as a function of I_R and I_L :

$$\mathbb{E}(I_1^2) = \mathbb{E}\left(\frac{(I_L + I_R)^2}{\sqrt{2}^2}\right) = \frac{\mathbb{E}(I_L^2) + \mathbb{E}(I_R^2) + 2\mathbb{E}(I_LI_R)}{2} = \frac{v + v + 2c}{2} = v + c$$
$$\mathbb{E}(I_2^2) = \mathbb{E}\left(\frac{(I_L - I_R)^2}{\sqrt{2}^2}\right) = \frac{\mathbb{E}(I_L^2) + \mathbb{E}(I_R^2) - 2\mathbb{E}(I_LI_R)}{2} = \frac{v + v - 2c}{2} = v - c$$
$$\mathbb{E}(I_1I_2) = \mathbb{E}\left(\frac{(I_L - I_R)(I_L + I_R)}{\sqrt{2}\sqrt{2}}\right) = \frac{\mathbb{E}(I_L^2) - \mathbb{E}(I_R^2)}{2} = \frac{v - v}{2} = 0$$

2 Hebbian learning algorithm

2.1 Standard Hebbian learning

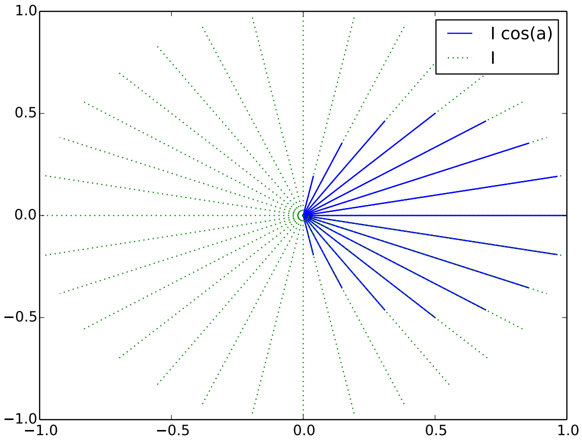
⑦ *Update of the weight vector*

- 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 a magnitude $\epsilon V(t)\|\vec{I}\|=\epsilon V(t)$ (under the assumption $\|\vec{I}\|=1$).

Moreover, the activity of the neuron is exactly the scalar product between $\vec{W(t)}$ and $\vec{I}(t)$: $V(t)=\vec{W}(t)\cdot\vec{I}(t)$. This scalar product can be interpreted with the cosine of the angle between both vectors : $V(t)=\|\vec{W}\|.\|\vec{I}\|\cos(\vec{W},\vec{I})=\|\vec{W}\|\cos(\alpha)$. Therefore, the update vector can be expressed as a function of α :

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

- Whatever the angle α, the norm of the weight vector $\|\vec{W}\|$ increases at each update. Indeed :
  - If $\vec{W},\vec{I}$ are oriented in 'similar direction', then $\alpha\in[-\pi/2,\pi/2]$ and $\cos(\vec{W},\vec{I})\geq 0$. The update vector is in the direction of $\vec{I}$, and consequently in the direction of $\vec{W}$ too.
  - If $\vec{W},\vec{I}$ are oriented in 'opposite direction', then $\alpha\in[\pi/2,3\pi/2]$ and $\cos(\vec{W},\vec{I})\leq 0$. The update vector is in the opposite direction of $\vec{I}$, and consequently still in the direction of $\vec{W}$.



⑧ *Update of along main axes*

If $\vec{W}$ is along one of the axes $\vec{e}_1$, $\vec{e}_2$, then through averaging, the update vector $\frac{\mathrm{d}\vec{W}}{\mathrm{d}t}$ will be parallel to $\vec{W}$.
Indeed, for instance with $\vec{W}=w\vec{e_1}$:

$$\begin{aligned}
\langle\Delta\vec{W}\rangle &= \langle V(t)\vec{I}(t)\rangle\\
&= \left\langle\vec{W}(t)\cdot\vec{I}(t)\times\vec{I}(t)\right\rangle\\
&= \left\langle w\,\underbrace{\vec{e}_1\cdot\vec{I}(t)}_{I_1}\times\underbrace{\vec{I}(t)}_{I_1\vec{e}_1+I_2\vec{e}_2}\right\rangle\\
&= w\,\langle I_1\times(I_1\vec{e}_1+I_2\vec{e}_2)\rangle\\
&= w\,\left(\langle I_1^2\rangle\vec{e}_1+\langle I_1I_2\rangle\vec{e}_2\right)\\
&= w\,((v+c)\times\vec{e}_1+0\times\vec{e}_2)\quad\text{question ⑥}\\
&= w(v+c)\vec{e}_1\\
&= (v+c)\vec{W}
\end{aligned}$$

Therefore, the axes $\vec{e}_1$, $\vec{e}_2$ are the *eigenvectors* of the dynamics :

- $\vec{e}_1$ is associated to the largest eigenvalue $v+c$,
- $\vec{e}_2$ is associated to the lowest eigenvalue $v-c$.

⑨ *Evolution of $\vec{W}$*

Method ① – In the basis of eigenvectors $\vec{e}_1, \vec{e}_2$

The dynamics can be expressed directly in the eigenvectors basis $\vec{e}_1, \vec{e}_2$ (question ⑧), writing $\vec{W} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$:

$$\begin{aligned}
\langle \Delta \vec{W} \rangle = \langle V(t) \vec{I}(t) \rangle &= \left\langle \begin{bmatrix} w_1 I_1^2 + w_2 I_1 I_2 \\ w_2 I_2^2 + w_1 I_1 I_2 \end{bmatrix} \right\rangle \\
&= \begin{bmatrix} w_1(v+c) + 0 \\ 0 + w_2(v-c) \end{bmatrix} \qquad \text{(question ⑧)} \\
&= \underbrace{\begin{pmatrix} v+c & 0 \\ 0 & v-c \end{pmatrix}}_{\mathbf{D}} \vec{W}
\end{aligned}$$

The system is already diagonalized, such that the evolution of the weights are governed by simple geometric sequences :

$$\begin{aligned}
\langle \Delta \vec{w_1} \rangle &= (v+c) w_1 \\
\langle \Delta \vec{w_2} \rangle &= (v-c) w_2
\end{aligned}$$

Method ② – In the initial basis $\vec{e}_L, \vec{e}_R$

The evolution of the weight vector can be expressed by a linear transformation in the basis $\vec{e}_L, \vec{e}_R$, writing $\vec{W} = \begin{bmatrix} w_L \\ w_R \end{bmatrix}$:

$$\begin{aligned}
\langle \Delta \vec{W} \rangle = \langle V(t) \vec{I}(t) \rangle &= \left\langle \vec{W}(t) \cdot \vec{I}(t) \times \vec{I}(t) \right\rangle \\
&= \left\langle (w_L I_L + w_R I_R) \begin{bmatrix} I_L \\ I_R \end{bmatrix} \right\rangle \\
&= \left\langle \begin{bmatrix} w_L I_L^2 + w_R I_L I_R \\ w_R I_R^2 + w_L I_L I_R \end{bmatrix} \right\rangle \\
&= \begin{bmatrix} w_L \langle I_L^2 \rangle + w_R \langle I_L I_R \rangle \\ w_R \langle I_R^2 \rangle + w_L \langle I_L I_R \rangle \end{bmatrix} \\
&= \begin{bmatrix} w_L v + w_R c \\ w_R v + w_L c \end{bmatrix} \\
&= \underbrace{\begin{pmatrix} v & c \\ c & v \end{pmatrix}}_{\mathbf{A}} \vec{W}
\end{aligned}$$

Projecting on each axis, this relation is equivalent to :

$$\begin{aligned}
\langle \Delta \vec{w_1} \rangle &= w_L v + w_R c \\
\langle \Delta \vec{w_2} \rangle &= w_R v + w_L c
\end{aligned}$$

Those equations are coupled. To solve them, the evolution of $\vec{W}$ can be expressed in the orthogonal axes defined by the eigenvectors of the matrix $\mathbf{A}$ (which are $\vec{e}_1$, $\vec{e}_2$ according to question ⑧, but in this method it is assumed that they have not been identified previously). The eigenvalues are found by the roots of the characteristic polynomial :

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

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 :

$$A \vec{e_1} = (v+c) \vec{e_1}$$

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

$$\begin{aligned}
\begin{pmatrix} v & c \\ c & v \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} v x_1 + c y_1 \\ c x_1 + v y_1 \end{pmatrix} &= \begin{pmatrix} (v+c) x_1 \\ (v+c) y_1 \end{pmatrix} \\
\Rightarrow x_1 = y_1 \Rightarrow \vec{e_1} &= \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}
\end{aligned}$$

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

$$\begin{pmatrix} v & c \\ c & v \end{pmatrix}\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} vx_2 + cy_2 \\ cx_2 + vy_2 \end{pmatrix} = \begin{pmatrix} (v-c)x_2 \\ (v-c)y_2 \end{pmatrix}$$

$$\Rightarrow x_2 = -y_2 \Rightarrow \vec{e_2} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

2.2 Improvements of Hebbian learning

(10) *Evolution of $\vec{W}$ with homeostasis*

- Is it *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 :

$$\begin{aligned}
\langle V(t)^2\rangle &= \langle(\vec{W}\cdot\vec{I})^2\rangle \\
&= \langle(w_1 I_1 + w_2 I_2)^2\rangle \\
&= w_1^2\langle I_1^2\rangle + w_2^2\langle I_2^2\rangle + w_1 w_2\langle I_1 I_2\rangle \\
&= w_1^2(v+c) + w_2^2(v-c) + 0 \\
\langle V(t)^2\rangle\vec{W} &= (w_1^2(v+c) + w_2^2(v-c))\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\
&= \begin{bmatrix} w_1^3(v+c) + w_1 w_2^2(v-c) \\ w_1^2 w_2(v+c) + w_2^3(v-c) \end{bmatrix}
\end{aligned}$$

- The evolution of the components of $\vec{W}$ obeys the following differential equations :

Note : terms in red stem from the standard hebbian rule (question (9)) and terms in blue correspond to the homeostatic term :

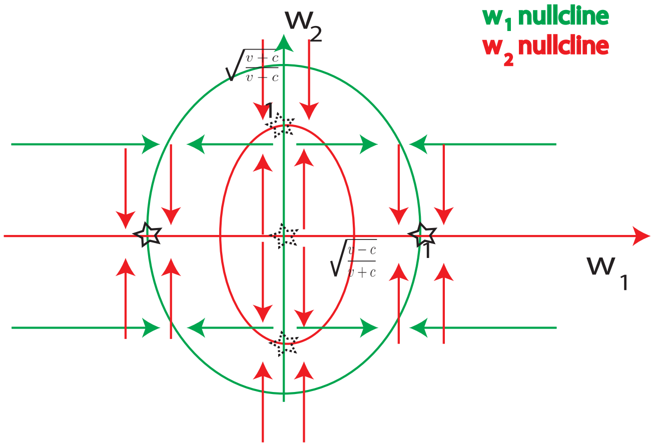
$$\begin{cases}\langle\Delta\vec{w_1}\rangle = (v+c)w_1 - (w_1^3(v+c) + w_1 w_2^2(v-c)) \\ \langle\Delta\vec{w_2}\rangle = (v-c)w_2 - (w_1^2 w_2(v+c) + w_2^3(v-c))\end{cases} \iff \begin{cases}\langle\Delta\vec{w_1}\rangle = w_1(v+c-w_1^2(v+c) - w_2^2(v-c)) \\ \langle\Delta\vec{w_2}\rangle = w_2(v-c-w_2^2(v-c) - w_1^2(v+c))\end{cases}$$

(11) *Nullclines and equilibria*

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

$$\begin{aligned}
\langle\Delta\vec{w_1}\rangle &= -(v+c)\,w_1\left(w_1^2 + \frac{v-c}{v+c}w_2^2 - 1\right) \\
\langle\Delta\vec{w_2}\rangle &= -(v-c)\,w_2\left(\frac{v+c}{v-c}w_1^2 + w_2^2 - 1\right)
\end{aligned}$$

- 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 respectively.
  - The w_1-nullcline contains the line $w_1 = 0$ and the ellipse $w_1^2 + \frac{v-c}{v+c}w_2^2 = 1$, which sets $a_1 = 1,\ b_1 = \sqrt{\frac{v+c}{v-c}} > 1$.
  - The w_2-nullcline contains the line $w_2 = 0$ and the ellipse $\frac{v+c}{v-c}w_1^2 + w_2^2 = 1$, which sets $a_2 = \sqrt{\frac{v-c}{v+c}} < 1,\ b_1 = 1$.



The arrows indicate the direction of evolution of the system in a given 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 the w_1 and w_2-nullclines. The only stable intersection points are :

$$\begin{cases} w_2 = 0 \\ w_1 = \pm 1 \end{cases}$$

⑫ *Interpretation 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.

⑬ *Evolution of $\vec{W}$ in competitive hebbian learning*

The second term of the learning rule can be expressed in the basis $(\vec{e}_1, \vec{e}_2)$, through the following relation (question ⑤) :

$$\begin{bmatrix} \frac{I_L+I_R}{2} \\ \frac{I_L+I_R}{2} \end{bmatrix} = \frac{I_L + I_R}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{\sqrt{2}I_1}{2}\sqrt{2}\vec{e_1} = I_1\vec{e_1}$$

Thereby :

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

- The evolution of the components of $\vec{W}$ obeys the following differential equations :

Note : terms in red stem from the standard hebbian rule (question ⑨) and terms in blue correspond to the competitive term :

$$\begin{cases} \langle \Delta \vec{w_1} \rangle = (v + c)w_1 - (v + c)w_1 \\ \langle \Delta \vec{w_2} \rangle = (v - c)w_2 \end{cases} \iff \begin{cases} \langle \Delta \vec{w_1} \rangle = 0 \\ \langle \Delta \vec{w_2} \rangle = (v - c)w_2 \end{cases} \iff \langle \Delta \vec{W} \rangle = \begin{pmatrix} 0 & 0 \\ 0 & v - c \end{pmatrix} \vec{W}$$

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

⑭ *Positive weights*

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.