1. One-step error probability in deterministic Hopfield model.

- Update rule: $S_i \leftarrow \text{sgn}\left(\sum_{j=1}^{N} w_{ij} S_j\right)$

- Weights: $\begin{cases} w_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)}, \text{ for } i \neq j \\ w_{ii} = 0 \end{cases}$

- Input patterns: $\underline{\xi}^{(\nu)}$; $\xi_i^{(\nu)}$ – bit $\underline{i}$ of input pattern $\underline{\xi}^{(\nu)}$; $\xi_i^{(\nu)} = +1$ or -1.

a) Condition for bit $\xi_i^{(\nu)}$ to be stable after a single step of asynchronous update?

Apply $\underline{\xi}^{(\nu)}$, obtain:

$$S_i = \text{sgn}\left[\sum_{j=1}^{N} w_{ij} \xi_j^{(\nu)}\right]$$

For stability of $\xi_i^{(\nu)}$ require: $\boxed{S_i \stackrel{!}{=} \xi_i^{(\nu)}}$ (*)

Rewrite the left-hand-side of Eq. (*):

$$S_i = \text{sgn}\left(\sum_{j=1}^{N} w_{ij} \xi_j^{(\nu)}\right) = \text{sgn}\left[\sum_{\substack{j=1 \\ j\neq i}}^{N} \left(\frac{1}{N}\sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)}\right) \xi_j^{(\nu)}\right] =$$

$$= \text{sgn}\left[\frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \xi_i^{(\nu)} \underbrace{\xi_j^{(\nu)} \xi_j^{(\nu)}}_{=1} + \frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{\substack{\mu=1 \\ \mu\neq\nu}}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\nu)}\right]$$

$$S_i = \text{sgn}\left[\frac{N-1}{N} \xi_i^{(\nu)} + \frac{1}{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{\substack{\mu=1 \\ \mu\neq\nu}}^{p} \xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\nu)}\right] \quad (\#)$$

Rewrite the right hand side of (#):

$$\text{RHS of } (\#) = \operatorname{sgn}\Big[\xi_i^{(\nu)} - \frac{1}{N}\xi_i^{(\nu)} + \underbrace{\frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p} \xi_i^{(\mu)}\xi_j^{(\mu)}\xi_j^{(\nu)}}_{\text{"cross-talk term"}}\Big]$$

Stability condition:

$$(**)\quad \boxed{\xi_i^{(\nu)} \overset{!}{=} \operatorname{sgn}\Big[\Big(1-\frac{1}{N}\Big)\xi_i^{(\nu)} + \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p} \xi_i^{(\mu)}\xi_j^{(\mu)}\xi_j^{(\nu)}\Big]}$$

Stability condition satisfied when:

$$\boxed{\Big|-\frac{1}{N}\xi_i^{(\nu)} + \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p} \xi_i^{(\mu)}\xi_j^{(\mu)}\xi_j^{(\nu)}\Big| < 1}$$

Alternatively, one can define $C_i^{(\nu)}$ as follows:

$$C_i^{(\nu)} = \frac{1}{N} - \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p} \xi_i^{(\mu)}\xi_j^{(\mu)}\xi_j^{(\nu)}\xi_i^{(\nu)}$$

(= cross-talk term $\times\,(-\xi_i^{(\nu)})$)

Multiply (**) by $(-\xi_i^{(\nu)})$ and rewrite the stability condition (**) as follows:

$$\boxed{-1 \overset{!}{=} \operatorname{sgn}\big(-1 + C_i^{(\nu)}\big)}$$

This condition is satisfied for $\boxed{C_i^{(\nu)} < 1}$.

Note: no limits were taken so far. In the limit of $N \gg 1$, $C_i^{(\nu)}$ is

$$C_i^{(\nu)} \approx -\frac{1}{N}\sum_{j=1}^{N}\sum_{\mu=1}^{p} \xi_i^{(\mu)}\xi_j^{(\mu)}\xi_j^{(\nu)}\xi_i^{(\nu)}, \quad \text{for } N \gg 1$$

b) Random patterns : $\zeta_i^{(\mu)} = \begin{cases} +1, & \text{with prob. } \frac{1}{2}, \\ -1, & \text{with prob } \frac{1}{2}. \end{cases}$

Bit $\zeta_i^{(\nu)}$ is stable after a single step of asynchronous update if $C_i^{(\nu)} < 1$ (task a).
Therefore, the probability (Perror) that $\zeta_i^{(\nu)}$ is unstable is:

$$\boxed{P_{error} = \text{Prob}\left(C_i^{(\nu)} > 1 \right)}$$

To evaluate Perror, consider $C_i^{(\nu)}$:

$$C_i^{(\nu)} = \frac{1}{N} - \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{p} \zeta_i^{(\mu)} \zeta_j^{(\mu)} \zeta_i^{(\nu)} \zeta_j^{(\nu)} \Rightarrow$$

$$C_i^{(\nu)} \overset{N \gg 1}{\approx} -\frac{1}{N} \underbrace{\sum_{k=1}^{(p-1)(N-1)}}_{\boxed{(p-1)(N-1) \text{ terms}}} \text{random variables } (x_k) \text{ with } \pm 1$$

Since we assume $p \gg 1$ and $N \gg 1$, we can use the Central limit theorem (patterns are random!)
Variables x_k have the mean $\underline{0}$, and variance $\sigma_k^2 = 1$.
It follows that $C_i^{(\nu)}$ has the following properties:

- $C_i^{(\nu)}$ is approximately Gaussian distributed,
- the mean of $C_i^{(\nu)}$ is equal to $\underline{0}$ (since the mean of the random variables x_k is $\underline{0}$)
- the variance σ^2 of $C_i^{(\nu)}$ is:

$$\sigma^2 = \frac{1}{N^2} \cdot (N-1)(p-1)\sigma_k^2 \approx \frac{p}{N}$$

$$\Rightarrow \sigma^2 \approx \frac{p}{N} \quad (\text{since } p \gg 1,\ N \gg 1)$$

It follows that

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-y^2}dy$$

$$P_{error} = \underbrace{\int_1^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}}_{\text{Gaussian distribution}} dx = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{1}{\sqrt{2}\,\sigma}\right)\right]$$

$$\Rightarrow P_{error} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{1}{\sqrt{2\frac{p}{N}}}\right)\right]$$

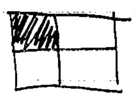
$$\boxed{P_{error} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{\frac{N}{2p}}\right)\right]}$$

② Hopfield model: recognition of one pattern.

Stored pattern: $\underline{\xi}^{(1)} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$

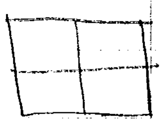
weight matrix: $\underset{\sim}{w} = \frac{1}{N}\underline{\xi}^{(1)}\underline{\xi}^{(1)T} = \frac{1}{4}\begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix}\begin{pmatrix} 1 & -1 & -1 & -1 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}$

- Feeding in the 2^4 possible patterns:

1) 

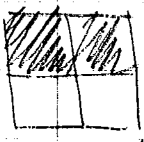
$\underline{S_0} = \underline{\xi}^{(1)} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$

$\rightarrow \underline{S_1} = \mathrm{sgn}(\underset{\sim}{w}\,\underline{\xi}^{(1)}) = \frac{1}{4}\underline{\xi}^{(1)}\underline{\xi}^{(1)T}\underline{\xi}^{(1)} = \frac{1}{4}\cdot 4\,\underline{\xi}^{(1)} = \underline{\xi}^{(1)}$

2) 

$\underline{S_0} = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$

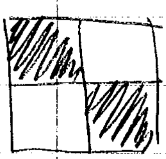
$\rightarrow \underline{S_1} = \mathrm{sgn}(\underset{\sim}{w}\,\underline{S_0}) = \frac{1}{4}\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \underline{\xi}^{(1)}$

3) 

$\underline{S}_0 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$

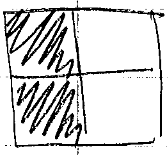
$$\rightarrow \underline{S}_1 = \operatorname{sgn}(\underset{\sim}{w}, \underline{S}_0) = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}\right] =$$

$$= \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \underline{y}^{(1)}$$

4) 

$\underline{S}_0 = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$

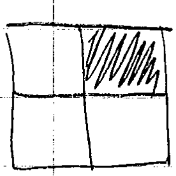
$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}\right] = \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \underline{y}^{(1)}$$

5) 

$\underline{S}_0 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$

$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}\right] = \begin{pmatrix} 1 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$

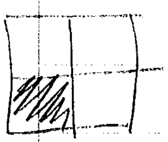
$$= \underline{y}^{(1)}$$

6) 

$\underline{S}_0 = \begin{pmatrix} -1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$

$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ -1 \\ -1 \end{pmatrix}\right] = \underline{\underline{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}}}$$

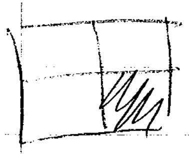
Orthogonal pattern to the stored pattern. The network doesnot restore the stored pattern. In fact, it retrieves zero vector; failure of the network performance.

7) 

$\underline{S}_0 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$

$$\underline{S}_1 = \mathrm{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ -1 \\ 1 \\ -1 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

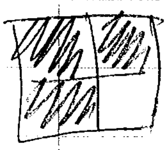
Same as case 6 : orthogonal pattern.

8) 

$\underline{S}_0 = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$

$$\underline{S}_1 = \mathrm{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ -1 \\ -1 \\ 1 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

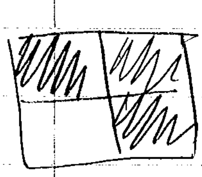
Same as cases 6-7.

9) 

$\underline{S}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix}$

$$\underline{S}_1 = \mathrm{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

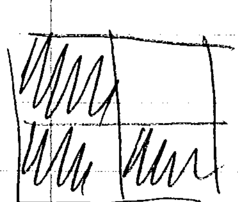
Same as cases 6-8.

10) 

$\underline{S}_0 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ 1 \end{pmatrix}$

$$\underline{S}_1 = \mathrm{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ -1 \\ 1 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

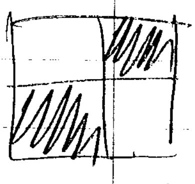
Same as cases 6-9.

11) 

$\underline{S}_0 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ 1 \end{pmatrix}$

$$\underline{S}_1 = \mathrm{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 1 \\ 1 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

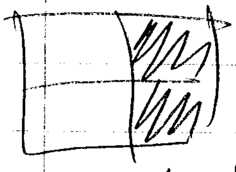
Same as cases 6-10.

12) 

$\underline{S}_0 = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \end{pmatrix}$

$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \end{pmatrix}\right] = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

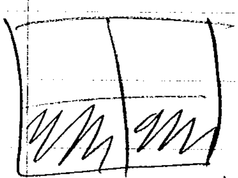
$$\boxed{\underline{S}_1 = -\underline{y}^{(1)}}$$

13) 

$\underline{S}_0 = \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix}$

$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ +1 \\ -1 \\ +1 \end{pmatrix}\right] = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

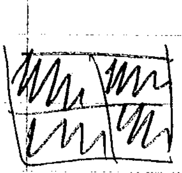
$$\boxed{\underline{S}_1 = -\underline{y}^{(1)}}$$

14) 

$\underline{S}_0 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}$

$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}\right] =$$

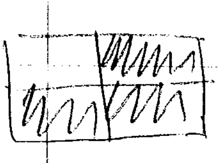
$$= \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = -\underline{y}^{(1)}$$

15) 

$\underline{S}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}\right] = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \Rightarrow$$

$$\boxed{\underline{S}_1 = -\underline{y}^{(1)}}$$

16) 

$\underline{S}_0 = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

$$\underline{S}_1 = \operatorname{sgn}\left[\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}\right] = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

In summary: 1 In the first 5 cases, the network retreives the stored pattern.
Note: in cases 2, 3, 4, 5, the pattern that was fed had only one distorted bit in comparison to the stored pattern.
Case 1: fed pattern = stored pattern.

2 In cases when more than 2 bits are distorted, the network retreives the inverted version of the stored pattern (cases 12 - 16)

3 When exactly $\frac{N}{2} = 2$ bits are distorted, the network fails (is unable to deal with patterns orthogonal to the stored pattern (due to Hebb's rule).

cases 6-11

13 Back-propagation (two hidden layers)

- Two hidden layers.
- Input patterns $\xi^{(\mu)} = (\xi_1, \xi_2, \dots, \xi_N)^T$
- Target output $\zeta_1^{(\mu)}$
- Network output $O_1^{(\mu)}$

- First hidden layer: $V_j^{(1,\mu)} = g(b_j^{(1,\mu)})$, $b_j^{(1,\mu)} = \sum_i w_{ji}^{(1)} \xi_i^{(\mu)} - \theta_j^{(1)}$

- Second hidden layer: $V_k^{(2,\mu)} = g(b_k^{(2,\mu)})$, $b_k^{(2,\mu)} = \sum_j w_{kj}^{(2)} V_j^{(1,\mu)} - \theta_k^{(2)}$

- Output layer: $O_1^{(\mu)} = g(b_1^{(\mu)})$, $b_1^{(\mu)} = \sum_k W_{1k} V_k^{(2,\mu)} - \Theta_1$

- Energy function: $H = \frac{1}{2} \sum_{\mu} \left(y_1^{(\mu)} - O_1^{(\mu)} \right)^2$

- Gradient descent: find the parameters that minimise $\underline{\underline{H}}$.

- Start from the output layer:

$$\delta W_{1k} = -\eta \frac{\partial H}{\partial W_{1k}} = -\eta \frac{\partial}{\partial W_{1k}} \left\{ \frac{1}{2} \sum_{\mu} \left[y_1^{(\mu)} - g(b_1^{(\mu)}) \right]^2 \right\} =$$

$$= -\eta \left[\sum_{\mu} \left[y_1^{(\mu)} - \underbrace{g(b_1^{(\mu)})}_{O_1^{(\mu)}} \right] \right] \left(- \frac{\partial g(b_1^{(\mu)})}{\partial W_{1k}} \right) =$$

$$= \eta \left[\sum_{\mu} \left[y_1^{(\mu)} - O_1^{(\mu)} \right] \cdot \frac{\partial g(b_1^{(\mu)})}{\partial W_{1k}} \right.$$

$$\frac{\partial g(b_1^{(\mu)})}{\partial W_{1k}} = \frac{\partial}{\partial W_{1k}} \left[g\left(\sum_{\ell} W_{1\ell} V_\ell^{(2,\mu)} - \Theta_1 \right) \right] =$$

$$= \underline{\underline{g'(b_1^{(\mu)}) \cdot V_k^{(2,\mu)}}} \qquad // \text{ Since } \frac{\partial W_{1\ell}}{\partial W_{1k}} = \delta_{\ell k}$$

$$\Rightarrow \boxed{\delta W_{1k} = \eta \sum_{\mu} \left[y_1^{(\mu)} - O_1^{(\mu)} \right] \cdot g'(b_1^{(\mu)}) \cdot V_k^{(2,\mu)}} \equiv \eta \sum_{\mu} \delta_1^{(3,\mu)} V_k^{(2,\mu)}$$

$$\delta \Theta_1 = -\eta \frac{\partial H}{\partial \Theta_1} = -\eta \frac{\partial}{\partial \Theta_1} \left\{ \frac{1}{2} \sum_{\mu} \left[y_1^{(\mu)} - g(b_1^{(\mu)}) \right]^2 \right\} =$$

$$= -\eta \sum_{\mu} \left(y_1^{(\mu)} - O_1^{(\mu)} \right) \cdot \left(- \frac{\partial g(b_1^{(\mu)})}{\partial \Theta_1} \right) =$$

$$= \eta \sum_{\mu} \left(y_1^{(\mu)} - O_1^{(\mu)} \right) \cdot g'(b_1^{(\mu)}) \cdot (-1)$$

$$\Rightarrow \boxed{\delta \Theta_1 = -\eta \sum_{\mu} \left(y_1^{(\mu)} - O_1^{(\mu)} \right) \cdot g'(b_1^{(\mu)})} \equiv -\eta \sum_{\mu} \delta_1^{(3,\mu)}$$

$$\boxed{\delta_1^{(3,\mu)} \equiv \left(y_1^{(\mu)} - O_1^{(\mu)} \right) g'(b_1^{(\mu)})}$$

- Second hidden layer

$$\delta w_{kj}^{(2)} = -\eta \frac{\partial H}{\partial w_{kj}^{(2)}} = -\eta \frac{\partial}{\partial w_{kj}^{(2)}} \left\{ \frac{1}{2} \sum_{\mu} \left(\varsigma_1^{(\mu)} - O_1^{(\mu)} \right)^2 \right\} =$$

$$= \eta \sum_{\mu} \left(\varsigma_1^{(\mu)} - O_1^{(\mu)} \right) \frac{\partial O_1^{(\mu)}}{\partial w_{kj}^{(2)}}$$

$$O_1^{(\mu)} = g(b_1^{(\mu)}) = g\left[\sum_{\ell} W_{1\ell} V_\ell^{(2,\mu)} - \Theta_1 \right] =$$

$$= g\left[\sum_{\ell} W_{1\ell}\, g(b_\ell^{(2,\mu)}) - \Theta_1 \right] =$$

$$= g\left[\sum_{\ell} W_{1\ell}\, g\left(\sum_{s} w_{\ell s}^{(2)} V_s^{(1,\mu)} - \theta_\ell^{(2)} \right) - \Theta_1 \right]$$

$$\Rightarrow \frac{\partial O_1^{(\mu)}}{\partial w_{kj}^{(2)}} = g'(b_1^{(\mu)}) \cdot \frac{\partial}{\partial w_{kj}^{(2)}} \left[\sum_{\ell} W_{1\ell}\, g\Big(\underbrace{\sum_{s} w_{\ell s}^{(2)} V_s^{(1,\mu)} - \theta_\ell^{(2)}}_{=b_\ell^{(2,\mu)}} \Big) - \Theta_1 \right]$$

$$= g'(b_1^{(\mu)}) \cdot \sum_{\ell} W_{1\ell}\, g'(b_\ell^{(2,\mu)}) \cdot \underbrace{\frac{\partial b_\ell^{(2,\mu)}}{\partial w_{kj}^{(2)}}}_{= \sum_s V_s^{(1,\mu)} \delta_{k\ell} \delta_{js}} =$$

$$= g'(b_1^{(\mu)}) \cdot W_{1k}\, g'(b_k^{(2,\mu)}) \cdot V_j^{(1,\mu)}$$

$$\Rightarrow \delta w_{kj}^{(2)} = \eta \sum_{\mu} \underbrace{\left(\varsigma_1^{(\mu)} - O_1^{(\mu)} \right) g'(b_1^{(\mu)})}_{\delta_1^{(3,\mu)}} \cdot W_{1k}\, g'(b_k^{(2,\mu)}) V_j^{(1,\mu)}$$

$$\delta w_{kj}^{(2)} = \eta \sum_{\mu} \underbrace{\delta_1^{(3,\mu)} W_{1k}\, g'(b_k^{(2,\mu)})}_{\delta_k^{(2,\mu)}} V_j^{(1,\mu)}$$

$$\boxed{\delta w_{kj}^{(2)} = \eta \sum_{\mu} \delta_k^{(2,\mu)} V_j^{(1,\mu)}}$$

Thresholds $\theta_k^{(2)}$:

$$\delta\theta_k^{(2)} = -\eta \frac{\partial H}{\partial \theta_k^{(2)}} = \eta \sum_\mu \left(\varsigma_1^{(\mu)} - O_1^{(\mu)}\right) \frac{\partial O_1^{(\mu)}}{\partial \theta_k^{(2)}}$$

$$\frac{\partial O_1^{(\mu)}}{\partial \theta_k^{(2)}} \overset{\text{from previous page}}{=} g'(b_1^{(\mu)}) \frac{\partial}{\partial \theta_k^{(2)}} \left[\sum_\ell W_{1\ell}\, g\left(\sum_s w_{\ell s}^{(2)} V_s^{(1,\mu)} - \theta_\ell^{(2)}\right) - \Theta_1 \right]$$

$$= g'(b_1^{(\mu)}) \sum_\ell W_{1\ell}\, g'(b_\ell^{(2,\mu)})\, (-1)\, \delta_{\ell k}$$

$$= -g'(b_1^{(\mu)}) \cdot W_{1k}\, g'(b_k^{(2,\mu)})$$

$$\Rightarrow \delta\theta_k^{(2)} = -\eta \sum_\mu \underbrace{\left(\varsigma_1^{(\mu)} - O_1^{(\mu)}\right) g'(b_1^{(\mu)})}_{\delta_1^{(3,\mu)}} W_{1k}\, g'(b_k^{(2,\mu)})$$

$$= -\eta \sum_\mu \underbrace{\delta_1^{(3,\mu)} W_{1k}\, g'(b_k^{(2,\mu)})}_{\delta_k^{(2,\mu)}}$$

$$\boxed{\delta\theta_k^{(2)} = -\eta \sum_\mu \delta_k^{(2,\mu)}}$$

For the first hidden layer we should proceed as above. Alternatively, we note that δ's for the 3rd and 2nd layer obey the following relation:

$$\delta_k^{(2,\mu)} = \delta_1^{(3,\mu)} W_{1k}\, g'(b_k^{(2,\mu)})$$

We can use this to find the δ's for the first hidden layer:

$$\delta_j^{(1,\mu)} = \sum_k \delta_k^{(2,\mu)} w_{kj}^{(2)} g'(b_j^{(1,\mu)})$$

The update formulae are, therefore, as follows:

Output layer: $$\delta W_{1k} = \eta \sum_\mu \delta_1^{(3,\mu)} V_k^{(2,\mu)}$$

$$\delta \Theta_1 = -\eta \sum_\mu \delta_1^{(3,\mu)}$$

Second hidden layer: $$\delta w_{kj}^{(2)} = \eta \sum_\mu \delta_k^{(2,\mu)} V_j^{(1,\mu)}$$

$$\delta \theta_k^{(2)} = -\eta \sum_\mu \delta_k^{(2,\mu)}$$

First hidden layer: $$\delta w_{ji}^{(1)} = \eta \sum_\mu \delta_j^{(1,\mu)} \xi_i^{(\mu)}$$

$$\delta \theta_j^{(1)} = -\eta \sum_\mu \delta_j^{(1,\mu)}$$

Summation over μ only for batch mode. Otherwise: no summation!

Here we have the following:

$$\delta_1^{(3,\mu)} = (\zeta_1^{(\mu)} - O_1^{(\mu)}) g'(b_1^{(\mu)}), \quad b_1^{(\mu)} = \sum_k W_{1k} V_k - \Theta_1$$

$$\delta_k^{(2,\mu)} = \delta_1^{(3,\mu)} W_{1k} g'(b_k^{(2,\mu)}), \quad b_k^{(2,\mu)} = \sum_j w_{kj}^{(2)} V_j^{(1,\mu)} - \theta_k^{(2)}$$

$$\delta_j^{(1,\mu)} = \sum_k \delta_k^{(2,\mu)} w_{kj}^{(2)} g'(b_j^{(1,\mu)}), \quad b_j^{(1,\mu)} = \sum_i w_{ji}^{(1)} \xi_i^{(\mu)} - \theta_j^{(1)}$$

(4) Backpropagation II — discussion of the implementation of the algorithm above. Explain how you program backpropagation.

17) Oja's rule: Output $\zeta = \sum_{j=1}^{N} w_j \xi_j = \underline{w}^T \underline{\xi}$

$\delta w_j = \eta \zeta (\xi_j - \zeta w_j)$

a) prove that $\underline{w}^*$ maximises $\langle \zeta^2 \rangle$ using that

$|\underline{w}^*|^2 = 1$ and $\underline{w}^*$ is the leading eigenvector of $\mathbb{C}$, with elements $C_{ij} = \langle \xi_i \xi_j \rangle$.

$$\langle \zeta^2 \rangle = \langle (\underline{w}^T \underline{\xi})(\underline{\xi}^T \underline{w}) \rangle = \langle \underline{w}^T \mathbb{C} \underline{w} \rangle$$

For $\underline{w} = \underline{w}^*$, find $\langle \zeta^2 \rangle_{\underline{w}^*} = \langle \underline{w}^{*T} \underbrace{\mathbb{C} \underline{w}^*}_{\substack{\lambda_{max} \underline{w}^* \\ \text{(from ii)}}} \rangle = \lambda_{max} \langle \underbrace{\underline{w}^{*T} \underline{w}^*}_{\substack{=1 \\ \text{(from i)}}} \rangle$

$\Rightarrow \boxed{\langle \zeta^2 \rangle_{\underline{w}^*} = \lambda_{max}}$, where λ_{max} is the maximum eigenvalue of $\mathbb{C}$.

Since $\mathbb{C}$ is symmetric $(\langle \xi_i \xi_j \rangle = \langle \xi_j \xi_i \rangle)$ it has real eigenvalues λ_α and its eigenvectors $\underline{u}_\alpha$ are orthogonal:

$\underline{u}_\alpha \underline{u}_\beta = \delta_{\alpha\beta}$, where $\delta_{\alpha\beta} = \begin{cases} 1, & \text{for } \alpha = \beta \\ 0, & \text{otherwise} \end{cases}$

Furthermore, all eigenvalues of $\mathbb{C}$ are positive, since

$$\lambda_\alpha = \underline{u}_\alpha^T \mathbb{C} \underline{u}_\alpha = \underline{u}_\alpha^T \langle \underline{\xi} \underline{\xi}^T \rangle \underline{u}_\alpha = \langle \underline{u}_\alpha^T \underline{\xi} \, \underline{\xi}^T \underline{u}_\alpha \rangle =$$

$$= \langle |\underline{u}_\alpha^T \underline{\xi}|^2 \rangle \geq 0$$

For any unit vector $\underline{w} = \sum_\alpha k_\alpha \underline{u}_\alpha$ that can be represented as a linear combination of the eigenvectors $\underline{u}_\alpha$ with coefficients k_α (assuming that $|\underline{w}|^2 = 1$) we find

$$\langle \zeta^2 \rangle_{\underline{w}} = \langle \Big(\sum_\alpha k_\alpha \underline{u}_\alpha\Big)^T \mathbb{C} \Big(\sum_\beta k_\beta \underline{u}_\beta\Big) \rangle = \langle \sum_\alpha (k_\alpha \underline{u}_\alpha)^T \Big(\sum_\beta k_\beta \lambda_\beta \underline{u}_\beta\Big) \rangle =$$

$$= \langle \sum_{\alpha,\beta} k_\alpha k_\beta \lambda_\beta \underbrace{\underline{u}_\alpha^T \underline{u}_\beta}_{\delta_{\alpha\beta}} \rangle = \langle \sum_\alpha (k_\alpha)^2 \lambda_\alpha \rangle \leq \lambda_{max} \langle \sum_\alpha |k_\alpha|^2 \rangle$$

From $|\underline{w}|^2 = 1$, we find $\sum_\alpha (K_\alpha)^2 = 1$

Therefore: $\langle \zeta^2 \rangle_{\underline{w}} \le \lambda_{max} \langle \sum_\alpha |K_\alpha|^2 \rangle = \lambda_{max}$

$$\Downarrow$$

$\boxed{\langle \zeta^2 \rangle_{\underline{w}} \le \lambda_{max}}$ and $\langle \zeta^2 \rangle_{\underline{w}^*} = \lambda_{max}$

This shows that $\langle \zeta^2 \rangle_{\underline{w}^*}$ is maximal in comparison to $\langle \zeta^2 \rangle_{\underline{w}}$ evaluated for any other $\underline{w}$ such that $|\underline{w}|^2 = 1$.

b) Assume that $\underline{w}^*$ is a steady state. In other words:

$$\langle \delta \underline{w} \rangle_{\underline{w}^*} = 0$$

$$\Rightarrow \langle \eta \zeta (\underline{\xi} - \zeta \underline{w}) \rangle_{\underline{w}^*} = 0$$

$$\Rightarrow \langle \underline{w}^{*T} \underline{\xi} (\underline{\xi} - \underline{w}^{*T} \underline{\xi} \underline{w}^*) \rangle = 0$$

/ $(\underline{w}^{*T} \underline{\xi}) \underline{\xi} = \underline{\xi} (\underline{w}^{*T} \underline{\xi}) = \underbrace{\underline{\xi} \underline{\xi}^T}_{\mathbb{C}} \underline{w}^*$

$$\langle \underbrace{\underline{\xi} \underline{\xi}^T}_{} \underline{w}^* - \underline{w}^{*T} \underbrace{\underline{\xi} \underline{\xi}^T}_{\mathbb{C}} \underline{w}^* \underline{w}^* \rangle = 0$$

$$\mathbb{C} \underline{w}^* - \underbrace{(\underline{w}^{*T} \mathbb{C} \underline{w}^*)}_{\text{scalar; let's call it } \underline{\lambda}} \underline{w}^* = 0$$

$(*\#) \Rightarrow \boxed{\mathbb{C} \underline{w}^* = \lambda \underline{w}^*} \Rightarrow$ Thus, $\underline{w}^*$ is an eigenvector of $\mathbb{C}$, with eigenvalue $\boxed{\lambda = \underline{w}^{*T} \mathbb{C} \underline{w}^*}$

Norm of $\underline{w}^*$ (property $\underline{i}$)

$$\lambda = \underbrace{\underline{w}^{*T} \mathbb{C} \underline{w}^*}_{\text{from } (*\#)} = \underline{w}^{*T} \lambda \underline{w}^* = \lambda \underline{w}^{*T} \underline{w}^* = \lambda |\underline{w}^*|^2$$

$\Rightarrow \boxed{|\underline{w}^*|^2 = 1}$ Shown (i)

Now we must show that $\underline{w}^*$ has the maximum eigenvalue $\underline{\lambda_{max}}$. Note: In order for the network to <u>converge</u> to a steady state, this steady state needs to be <u>stable</u>. Otherwise, the network would not converge to it.

Therefore, check the stability of $\underline{w}^*$.

Evaluate $\langle \delta \underline{w} \rangle$ at $\underline{w} = \underline{w}^* + \underline{\varepsilon}$, where $|\underline{\varepsilon}|$ is small.

$$\langle \delta(\underline{w}^* + \underline{\varepsilon}) \rangle = \eta \left\langle (\underline{w}^* + \underline{\varepsilon})^T \underline{\xi} \left[\underline{\xi} - (\underline{w}^* + \underline{\varepsilon})^T \underline{\xi} (\underline{w}^* + \underline{\varepsilon}) \right] \right\rangle$$

up to linear order in $\underline{\varepsilon}$

$$\approx \eta \Big[\underbrace{\langle \underline{w}^{*T} \underline{\xi} (\underline{\xi} - \underline{w}^{*T} \underline{\xi}\, \underline{w}^*) \rangle}_{= 0 \text{ because } \underline{w}^* \text{ is steady (previous page)}}$$

$$+ \underbrace{\langle \underline{\varepsilon}^T \underline{\xi}\, \underline{\xi} \rangle}_{= \langle \underline{\xi}\, \underline{\xi}^T \underline{\varepsilon} \rangle} - \langle \underline{\varepsilon}^T \underline{\xi} (\underline{w}^{*T} \underline{\xi}\, \underline{w}^*) \rangle$$

$$- \langle \underline{w}^{*T} \underline{\xi}\, \underline{w}^{*T} \underline{\xi}\, \underline{\varepsilon} \rangle$$

$$- \langle \underline{w}^{*T} \underline{\xi}\, \underline{\varepsilon}^T \underline{\xi}\, \underline{w}^* \rangle$$

say that $\underline{w}^* = \underline{u}_\alpha$, one of the eigenvectors.

$$\Rightarrow \langle \delta(\underline{w}^* + \underline{\varepsilon}) \rangle \approx \eta \Big[\langle \underline{\xi}\, \underline{\xi}^T \underline{\varepsilon} \rangle - \langle \underline{\varepsilon}^T \underbrace{\underline{\xi}\, \underline{\xi}^T}_{C} \underbrace{\underline{w}^*}_{=\underline{u}_\alpha} \underline{w}^* \rangle$$

$= \lambda_\alpha \underline{u}_\alpha$ $\qquad$ $\lambda_\alpha \underline{u}_\alpha$

$$- \langle \underline{w}^{*T} \underline{\xi}\, \underline{\xi}^T \underline{w}^* \underline{\varepsilon} \rangle - \langle \underline{w}^{*T} \underline{\xi}\, \underline{\xi}^T \underline{\varepsilon}\, \underline{w}^* \rangle \Big]$$

$$= \eta \Big[C \underline{\varepsilon} - \underline{\varepsilon}^T \lambda_\alpha \underline{u}_\alpha \underline{u}_\alpha$$

$$- \underbrace{\underline{u}_\alpha^T \lambda_\alpha \underline{u}_\alpha}_{= \lambda_\alpha} \underline{\varepsilon} - \lambda_\alpha \underbrace{\underline{u}_\alpha^T \underline{\varepsilon}}_{= \underline{\varepsilon}^T \underline{u}_\alpha} \underline{u}_\alpha \Big]$$

$$= \eta \left[C \underline{\varepsilon} - 2\lambda_\alpha (\underline{\varepsilon}^T \underline{u}_\alpha) \underline{u}_\alpha - \lambda_\alpha \underline{\varepsilon} \right)$$

Multiply both sides by $\underline{u}_\beta^T$ [illegible]

$$u_\beta^T \langle \delta(w^* + \varepsilon)\rangle = \eta\left(\overbrace{u_\beta^T C}^{=\lambda_\beta u_\beta^T}\varepsilon - 2\lambda_\alpha(\varepsilon^T u_\alpha)\, u_\beta^T u_\alpha - \lambda_\alpha u_\beta^T \varepsilon\right)$$

$$= \eta\,\underbrace{(\lambda_\beta - 2\lambda_\alpha\delta_{\alpha\beta} - \lambda_\alpha)}\, u_\beta^T \varepsilon$$

Recall: λ_α is the eigenvalue assigned to $\underline{w}^*$. Assume that this is not the maximal eigenvalue. In this case, thus, there will be at least one β with $\lambda_\beta > \lambda_\alpha$. In this case, it follows that an initially small fluctuation around $\underline{w}^*$ (denoted by $\underline{\varepsilon}$ above) will grow! This is because the right hand side of the equation above is, in this case, positive:

$$\lambda_\beta > \lambda_\alpha \Rightarrow (\lambda_\beta - 2\underbrace{\lambda_\alpha\delta_{\alpha\beta}}_{=0} - \lambda_\alpha) = \lambda_\beta - \lambda_\alpha > 0$$

Therefore, in this case $\underline{w}^*$ is not the weight vector to which the network converges.

What happens if λ_α is the maximum eigenvalue? From the above argument, find that $\underline{\varepsilon}$ will shrink in size in all directions $\underline{u}_\beta$ $(\beta \neq \alpha)$. What happens in the direction $\underline{u}_\alpha = \underline{w}^*$? In this direction $\underline{\varepsilon}$ also shrinks because the right-hand-side of the equation above is negative:

$$\lambda_\alpha - 2\lambda_\alpha - \lambda_\alpha = -2\lambda_\alpha < 0$$

Thus, we have shown that if the network converges to $\underline{w}^*$, then $\underline{w}^*$ is the leading eigenvector of C, and $|\underline{w}^*|^2 = 1$

c) Generalisation of Oja's rule for learning M principal components for zero-mean data

$$\delta w_{ij} = \eta \zeta_i \left(\xi_j - \sum_{k=1}^{M} \zeta_k w_{kj} \right)$$

where $\zeta_i = \sum_{j=1}^{N} w_{ij} \xi_j$.

When $M=1$, this rule reduces to the rule (5) in the exam text.
Weight decay (second term in the rule) assures that the weight vectors remain normalised.