CHALMERS, GÖTEBORGS UNIVERSITET

SOLUTIONS FOR EXAM for
ARTIFICIAL NEURAL NETWORKS

COURSE CODES: **FFR 135, FIM 720 GU, PhD**

Maximum score on this exam: 12 points.
Maximum score for homework problems: 12 points.
To pass the course it is necessary to score at least 5 points on this written exam.
CTH $\geq$14 passed; $\geq$17.5 grade 4; $\geq$22 grade 5,
GU $\geq$14 grade G; $\geq$ 20 grade VG.

---

1. Feature map.
Local fields of feature map of pattern $\boldsymbol{x}^{(1)}$:

$$\begin{bmatrix} 6 & 5 \\ 2 & 1 \\ 1 & 2 \\ 5 & 6 \end{bmatrix} . \tag{1}$$

Local fields of feature map of pattern $\boldsymbol{x}^{(2)}$:

$$\begin{bmatrix} 3 & 3 \\ 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix} . \tag{2}$$

The ReLU-activation function does not exert any effect, since all local fields are possible. The feature maps are therefore equal to the local fields above.
Max-pooling layer of pattern $\boldsymbol{x}^{(1)}$:

$$\begin{bmatrix} 6 \\ 2 \\ 6 \end{bmatrix} . \tag{3}$$

Max-pooling layer of pattern $\boldsymbol{x}^{(2)}$:

$$\begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix} . \tag{4}$$

With $W_k = -\delta_{k1}$ and $\Theta = -4$ we have

$$\sum_{k=1}^{3} W_k \begin{bmatrix} 6 \\ 2 \\ 6 \end{bmatrix}_k - \Theta = -2 \tag{5}$$

and

$$\sum_{k=1}^{3} W_k \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}_k - \Theta = 1. \tag{6}$$

Applying the Heaviside activation function results in the requested outputs.

2. Hopfield network with hidden units

Denote the the value of hidden neuron i after the update by h_i'. Suppose that the k^{th} hidden neuron change sign. We then have:

$$h_i' = h_i - 2h_i\delta_{ik}, \tag{7}$$

The energy after the update is

$$\begin{aligned}
H' &= -\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij} h_i' v_j && (8)\\
&= -\sum_{j=1}^{N} v_j \sum_{i=1}^{M} w_{ij}(h_i - 2h_i\delta_{ik}) && (9)\\
&= -\sum_{j=1}^{N} v_j \left[\sum_{i=1}^{M} w_{ij}h_i - 2\sum_{i=1}^{M} w_{ij}h_i\delta_{ik}\right] && (10)\\
&= -\sum_{j=1}^{N} v_j \left[\sum_{i=1}^{M} w_{ij}h_i - 2w_{kj}h_k\right] && (11)\\
&= -\sum_{j=1}^{N}\sum_{i=1}^{M} w_{ij}h_i v_j + 2h_k \sum_{j=1}^{N} w_{kj}v_j && (12)\\
&= H + 2h_k b_k^{(h)}. && (13)
\end{aligned}$$

If the k^{th} hidden neuron change sign, then $h_k b_k^{(h)} < 0$.

3. Backpropagation

Solution: The weight update rule for $L = 3$ reads,

$$\delta w_{pr}^{(3)} = -\eta \frac{\partial H}{\partial w_{pr}^{(3)}} \tag{14}$$

$$= \eta \sum_{i,\mu} (y_i^{(\mu)} - O_i^{(\mu)}) \frac{\partial V_i^{(3,\mu)}}{\partial w_{pr}^{(3)}} \tag{15}$$

$$= \eta \sum_{i,\mu} (y_i^{(\mu)} - O_i^{(\mu)}) g'(b_i^{(3,\mu)}) \sum_j \frac{\partial w_{ij}^{(3,\mu)}}{\partial w_{pr}^{(3)}} V_j^{(2,\mu)} \tag{16}$$

$$= \eta \sum_{i,\mu} (y_i^{(\mu)} - O_i^{(\mu)}) g'(b_i^{(3,\mu)}) \sum_j \delta_{ip} \delta_{jr} V_j^{(2,\mu)} \tag{17}$$

$$= \eta \sum_{\mu} (y_p^{(\mu)} - O_p^{(\mu)}) g'(b_p^{(3,\mu)}) V_r^{(2,\mu)} \tag{18}$$

$$= \eta \sum_{\mu} \Delta_p^{(3,\mu)} V_r^{(2,\mu)} \tag{19}$$

where, $\Delta_p^{(3,\mu)} = (y_p^{(\mu)} - O_p^{(\mu)}) g'(b_p^{(3,\mu)})$. Similarly,

$$\delta w_{pr}^{(2)} = -\eta \frac{\partial H}{\partial w_{pr}^{(2)}} \tag{20}$$

$$= \eta \sum_{i,\mu} (y_i^{(\mu)} - O_i^{(\mu)}) g'(b_i^{(3,\mu)}) w_{ip}^{(3)} g'(b_p^{(2,\mu)}) V_r^{(1,\mu)} \tag{21}$$

$$= \eta \sum_{i,\mu} \Delta_i^{(3,\mu)} w_{ip}^{(3)} g'(b_p^{(2,\mu)}) V_r^{(1,\mu)} \tag{22}$$

$$= \eta \sum_{\mu} \Delta_p^{(2,\mu)} V_r^{(1,\mu)} \tag{23}$$

where, $\Delta_p^{(2,\mu)} = \sum_{i,\mu} \Delta_i^{(3,\mu)} w_{ip}^{(3)} g'(b_p^{(2,\mu)})$. Similarly,

$$\delta w_{pr}^{(1)} = \eta \sum_{\mu} \Delta_p^{(1,\mu)} V_r^{(0,\mu)}, \tag{24}$$

where $\Delta_p^{(1,\mu)} = \sum_{i,\mu} \Delta_i^{(2,\mu)} w_{ip}^{(2)} g'(b_p^{(1,\mu)})$.

4. XNOR function.

$$3\boldsymbol{w} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \tag{25}$$

$$+ \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix} \tag{26}$$

$$= \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}. \tag{27}$$

The weight matrix is proportional to the identity matrix, and the network does not recognise the XNOR function.

$$H = -\frac{1}{2}\sum w_{ij}x_ix_j, \tag{28}$$

$$= -\frac{1}{2}\frac{4}{3}\sum \delta_{ij}x_ix_j \tag{29}$$

$$= -\frac{2}{3}\sum_i x_i^2 \tag{30}$$

$$= -\frac{2}{3}3 \tag{31}$$

$$= -2. \tag{32}$$

Thus the energy function is always constant and the network cannot learn. Further, even using the modified Hopfield rule in this case would not work, because the weight matrix would just be 0 and the energy would always be 0 as well.
Storing only 3 out of the 4 patterns makes the network linearly separable instead of linearly separable. Thus, now the network can recognise patterns.
5. Gradient descent and momentum
Consider the given energy function $\mathcal{H}$ as a function of weight w as shown in Fig. 2. Use the following gradient descent update rule,

$$\delta w_{n+1} = -\eta\frac{\partial \mathcal{H}}{\partial w} + \alpha\,\delta w_n. \tag{33}$$

Assume that the system is initially at point A, and that $\eta s = 1/2$. The slope of the segment AB in Fig. 2 is $-s$ and the slope of the segment BC is 0. The system starts at time step 1, and assume that $\delta w_0 = 0$.

1. Find the number of time steps required to travel from point A to point B for $\alpha = 0$.

2. Repeat the previous calculation for the case $\alpha = 1/2$, and graphically find the solution of the final equation you obtain.

3. Indicate the results of the previous two parts on the same graph. Which of the two cases: $\alpha = 0$ and $\alpha = 1/2$ converges faster?

4. What is the fate of the two systems $\alpha = 0$ and $\alpha = 1/2$ once they cross point B?

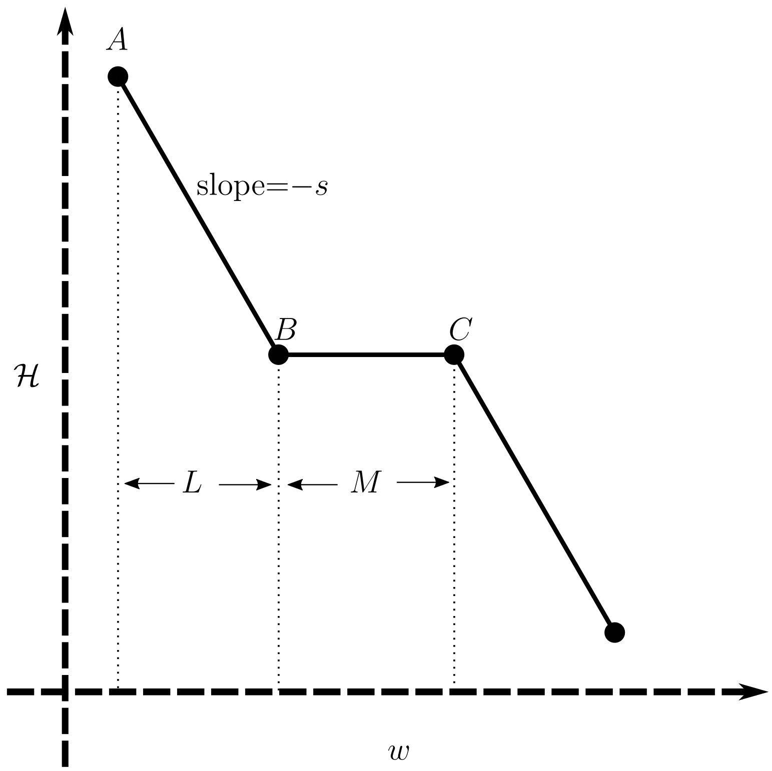

Figure 1: Energy as a function of weight for problem: Gradient descent and momentum.

Solution: 1 and 2: We calculate the total change in weight at time step $n, \Delta w_n = \sum_{i=1}^{n} \delta w_i$, equate Δw_n to L and solve for n. Proceed by solving for δw_n. Iterating the equation for δw we find,

$$\delta w_{i+1} = \sum_{j=0}^{i} \eta s \, \alpha^j + \alpha^{i+1} \delta w_0, \tag{34}$$

$$= \eta s \, \frac{1 - \alpha^{i+1}}{1 - \alpha}. \tag{35}$$

Next compute Δw_n,

$$\Delta w_n = \sum_{i=1}^{n} \delta w_i, \tag{36}$$

$$= \eta s \sum_{i=1}^{n} \frac{1-\alpha^{i+1}}{1-\alpha}, \tag{37}$$

$$= \frac{\eta s}{1-\alpha}\left(n - \alpha\frac{1-\alpha^n}{1-\alpha}\right). \tag{38}$$

Thus using $\eta s = 1/2$ we obtain, for $\alpha = 0$, $\Delta w_n(\alpha = 0) = n/2$, and for $\alpha = 1/2$, $\Delta w_n(\alpha = 1/2) = n - 1 + 2^{-m}$. Equating $\Delta w = L$ we obtain,

$$n_{\alpha=0} = 2L, \tag{39}$$

$$n_{\alpha=1/2} - 1 + 2^{-n_{\alpha=1/2}} = L. \tag{40}$$

graphing the above equations, we see that $n_{\alpha=1/2} < n_{\alpha=0}$, thus, $\alpha = 1/2$ converges faster.

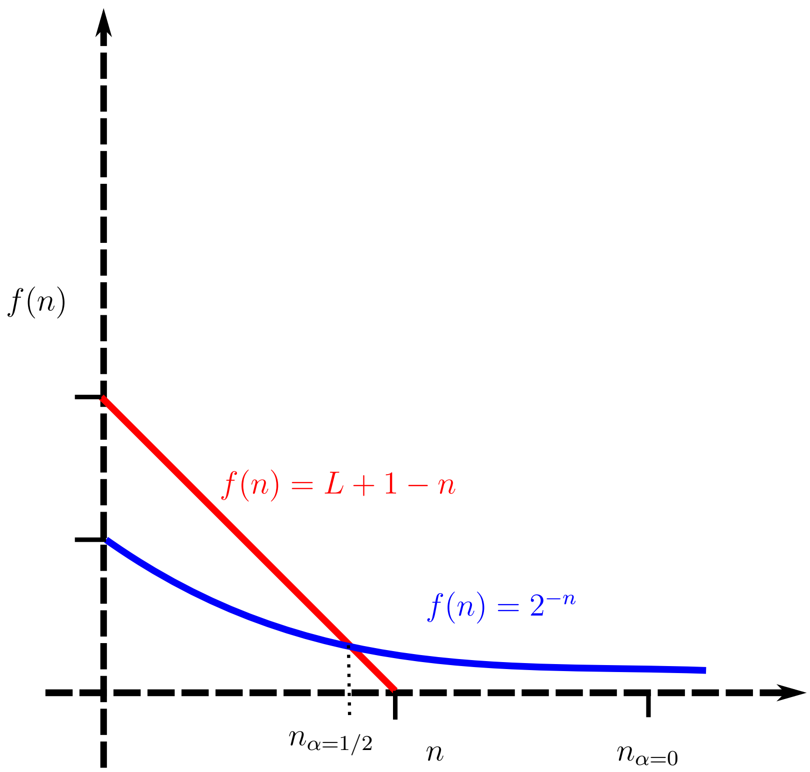

Figure 2: Graphical solution of problem : gradient descent and momentum.

After crossing point B, $\delta w(\alpha = 0) = 0$ so that this system stays stationary, however $\delta w_{\alpha=1/2} > 0$ so that this system keeps on moving.

6. Linear activation function

a)

$$
\begin{aligned}
\frac{\partial H}{\partial w_i} =& \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{\mu=1}^{p} \left(O^{(\mu)} - t^{(\mu)}\right)^2 && (41)\\
=& \sum_{\mu=1}^{p} \left(O^{(\mu)} - t^{(\mu)}\right) \frac{\partial O^{(\mu)}}{\partial w_i} && (42)\\
=& \sum_{\mu=1}^{p} \left(O^{(\mu)} - t^{(\mu)}\right) x_i^{(\mu)} && (43)\\
=& \sum_{\mu=1}^{p} \left(\sum_{j=1}^{N} w_j x_j^{(\mu)} - \theta - t^{(\mu)}\right) x_i^{(\mu)} && (44)\\
=& \sum_{j=1}^{N} \sum_{\mu=1}^{p} w_j x_j^{(\mu)} x_i^{\mu} - \theta \sum_{\mu=1}^{p} x_i^{\mu} - \sum_{\mu=1}^{p} t^{(\mu)} x_i^{\mu} && (45)\\
=& \sum_{j=1}^{N} w_j \sum_{\mu=1}^{p} x_j^{(\mu)} x_i^{\mu} - \theta \sum_{\mu=1}^{p} x_i^{\mu} - \sum_{\mu=1}^{p} t^{(\mu)} x_i^{\mu} && (46)\\
=& \sum_{j=1}^{N} w_j p G_{ji} - \theta p \beta_i - p \alpha_i && (47)\\
=& p \left(\sum_{j=1}^{N} G_{ij} w_j - \theta \beta_i - \alpha_i\right) && (48)\\
& && (49)
\end{aligned}
$$

$$
\frac{\partial H}{\partial w_i} = 0 \Rightarrow \mathbb{G} \boldsymbol{w} = \boldsymbol{\alpha} + \theta \boldsymbol{\beta} \tag{50}
$$

$$
\begin{aligned}
\frac{\partial H}{\partial \theta} =& \frac{\partial}{\partial \theta}\frac{1}{2}\sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)^2 && (51)\\
=& \sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)\frac{\partial O^{(\mu)}}{\partial \theta} && (52)\\
=& \sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)(-1) && (53)\\
=& \sum_{\mu=1}^{p}\left(\sum_{j=1}^{N} w_j x_j^{(\mu)} - \theta - t^{(\mu)}\right)(-1) && (54)\\
=& -\sum_{\mu=1}^{p}\sum_{j=1}^{N} w_j x_j^{(\mu)} + \sum_{\mu=1}^{p}\theta + \sum_{\mu=1}^{p} t^{(\mu)} && (55)\\
=& -p\sum_{j=1}^{N} w_j \beta_j + p\theta + pc && (56)\\
& && (57)
\end{aligned}
$$

$$
\frac{\partial H}{\partial \theta} = 0 \Rightarrow \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\beta} = \theta + \gamma. \tag{58}
$$

b) The first equation gives:

$$
\boldsymbol{w} = \mathbb{G}^{-1}\boldsymbol{\alpha} + \theta\mathbb{G}^{-1}\boldsymbol{\beta}. \tag{59}
$$

Insert into the second, and use that $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{w}$:

$$
\begin{aligned}
\boldsymbol{\beta}^{\mathsf{T}}\left[\mathbb{G}^{-1}\boldsymbol{\alpha} + \theta\mathbb{G}^{-1}\boldsymbol{\beta}\right] &= \theta + \gamma && (60)\\
\Rightarrow \boldsymbol{\beta}^{\mathsf{T}}\mathbb{G}^{-1}\boldsymbol{\alpha} + \theta\boldsymbol{\beta}^{\mathsf{T}}\mathbb{G}^{-1}\boldsymbol{\beta} &= \theta + \gamma && (61)\\
\Rightarrow \theta\left[\boldsymbol{\beta}^{\mathsf{T}}\mathbb{G}^{-1}\boldsymbol{\beta} - 1\right] &= \gamma - \boldsymbol{\beta}^{\mathsf{T}}\mathbb{G}^{-1}\boldsymbol{\alpha} && (62)\\
\Rightarrow \theta &= \frac{\gamma - \boldsymbol{\beta}^{\mathsf{T}}\mathbb{G}^{-1}\boldsymbol{\alpha}}{\boldsymbol{\beta}^{\mathsf{T}}\mathbb{G}^{-1}\boldsymbol{\beta} - 1}. && (63)
\end{aligned}
$$

c) The equation can be written as

$$
\boldsymbol{V}^{(\mu,\ell)} = \boldsymbol{W}\boldsymbol{V}^{(\mu,\ell-2)} - \boldsymbol{\Theta}, \tag{64}
$$

where

$$
\boldsymbol{W} = \boldsymbol{w}^{(\ell)}\boldsymbol{w}^{(\ell-1)}, \tag{65}
$$

and

$$
\boldsymbol{\Theta} = \boldsymbol{w}^{(l)}\boldsymbol{\theta}^{(\ell-1)} + \boldsymbol{\theta}^{(\ell)}. \tag{66}
$$

The two layers can therefore be collapsed into one single layer, and with a linear activation function in all layers the whole perceptron collapses into a simple perceptron with linear activation function. Such a perceptron can only solve linearly separable problems.