CHALMERS, GÖTEBORGS UNIVERSITET

SOLUTIONS FOR EXAM for
ARTIFICIAL NEURAL NETWORKS

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

1. Feature map. The two patterns $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ shown in Figure 1(a) are processed by a very simple convolutional network that has one convolution layer with one single 4×4 kernel with ReLU units, zero threshold, weights w_{ij} as given in Figure 1(b), and stride (1,1). The resulting feature map is fed into a 2×2 max-pooling layer with stride (1,1). Finally there is a fully connected output layer with one output unit $O^{(\mu)}$ with Heaviside activation function. For both patterns determine the resulting feature map and the output of the max-pooling layer. Determine weights W_k and a threshold Θ so that the network output is $O^{(1)} = 0$ for input pattern $\boldsymbol{x}^{(1)}$, and $O^{(2)} = 1$ for input pattern $\boldsymbol{x}^{(2)}$.

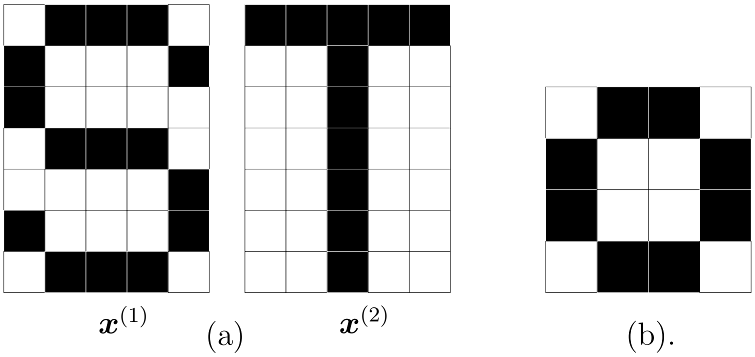

Figure 1: (a) Input patterns $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ with ±1 bits (□ corresponds to x_i=-1 and ■ to x_i=1). (b) Weights w_{ij} of a 4×4 kernel of a feature map. The weights are either -1 or 1 (□ corresponds to $w_{ij} = -1$ and ■ to $w_{ij} = 1$). (Question 1).

Solution: Input to feature map of pattern $\boldsymbol{x}^{(1)}$:

$$\begin{bmatrix} 8 & 6 \\ -2 & -6 \\ -6 & -2 \\ 6 & 8 \end{bmatrix}. \tag{1}$$

See Fig. 2 for an illustration of how to arrive at these numbers. Input to feature map of pattern $\boldsymbol{x}^{(2)}$:

$$\begin{bmatrix} -2 & -2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{2}$$

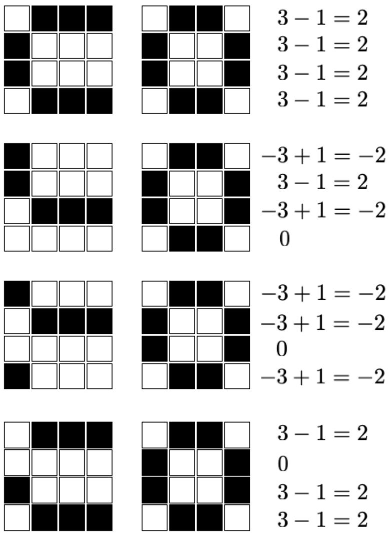

Figure 2: Illustrates calculation of input to feature map for the pattern S, Eq. (1), Question 1.

Feature map of pattern $\boldsymbol{x}^{(1)}$:

$$\begin{bmatrix} 8 & 6 \\ 0 & 0 \\ 0 & 0 \\ 6 & 8 \end{bmatrix}. \tag{3}$$

Feature map of pattern $\boldsymbol{x}^{(2)}$:

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

Max-pooling layer of pattern $\boldsymbol{x}^{(1)}$:

$$\begin{bmatrix} 8 \\ 0 \\ 8 \end{bmatrix}. \tag{5}$$

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

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6}$$

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

$$\sum_{k=1}^{3} W_k \begin{bmatrix} 8 \\ 0 \\ 8 \end{bmatrix}_k - \Theta = -4 \tag{7}$$

and

$$\sum_{k=1}^{3} W_k \begin{bmatrix}0\\0\\0\end{bmatrix}_k - \Theta = 4. \tag{8}$$

Applying the Heaviside activation function results in the requested outputs.

2. Hopfield network with hidden units A Hopfield network with hidden neurons can be used to learn a distribution of input patterns. Consider a Hopfield network with N visible neurons v_j and M hidden neurons h_i. The neurons are binary, with values -1 or $+1$. The network learns by updating the visible neurons according to

$$v_j \leftarrow \operatorname{sgn}\left[b_j^{(v)}\right] \quad \text{with} \quad b_j^{(v)} = \sum_{i=1}^{M} h_i w_{ij}, \tag{9}$$

and by updating the hidden neurons according to

$$h_i \leftarrow \operatorname{sgn}\left[b_i^{(h)}\right] \quad \text{with} \quad b_i^{(h)} = \sum_{j=1}^{N} w_{ij} v_j. \tag{10}$$

In Equations (9) and (10), w_{ij} are the elements of a $M \times N$ weight matrix. Furthermore, $\operatorname{sgn}[b]$ is the signum function, $\operatorname{sgn}[b] = -1$ if $b < 0$ and $+1$ otherwise. Show that the energy function

$$H = -\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij} h_i v_j \tag{11}$$

can not increase upon updating one of the hidden neurons according to equation (10).

Solution: Denote the the value of hidden neuron i after the update by h_i'. Suppose that the k^{th} hidden neuron changes sign. In this case:

$$h_i' = h_i - 2h_i\delta_{ik}, \tag{12}$$

The energy after the update is

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

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

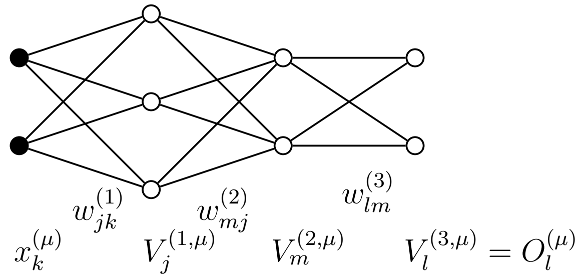

Figure 3: Network for Question 3.

3. Backpropagation

Assuming the energy function

$$
H = \frac{1}{2}\sum_{i,\mu}(y_i^{(\mu)} - O_i^{(\mu)})^2, \tag{13}
$$

derive the update rule for the weights $w_{ij}^{(L)}$ for $L = 1, 2$ and 3 for the network shown in Figure 3.

Solution: see course book.

4. XOR function. The Boolean XOR function takes two binary inputs. For the inputs $[-1,-1]$ and $[1,1]$ the function evaluates to -1, for the other two inputs it evaluates to $+1$. Encode the XOR function as weights w_{ij} in a Hopfield net with three neurons by storing the patterns $\boldsymbol{x}^{(1)} = [-1,-1,-1]$, $\boldsymbol{x}^{(2)} = [1,1,-1]$, $\boldsymbol{x}^{(3)} = [-1,1,1]$, and $\boldsymbol{x}^{(4)} = [1,-1,1]$ using Hebb's rule:

$$w_{ij} = \frac{1}{3}\sum_{\mu=1}^{4} x_i^{(\mu)} x_j^{(\mu)} \quad \text{where} \quad i,j = 1,\ldots,3. \tag{14}$$

The update rule for bit S_i is

$$S_i \leftarrow \operatorname{sgn}\left[\sum_{j=1}^{3} w_{ij} S_j\right], \tag{15}$$

where $\operatorname{sgn}[b]$ is the signum function, $\operatorname{sgn}[b] = -1$ if $b < 0$ and $+1$ otherwise. Feed the stored patterns to the net, and test whether they are stable under synchronous updating. Conclude with one or two sentences whether the network is useful for recognising the XOR function.

Solution:

$$\begin{aligned} 3\mathbb{W} &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}. \end{aligned} \tag{16}$$

The weight matrix is proportional to the identity matrix. Therefore the network reproduces all input patterns, it cannot single out the four patterns corresponding to the XOR function.

5. Gradient descent and momentum

Consider the given energy function $\mathcal{H}$ as a function of weight w as shown in Fig. 4. Use the following gradient descent update rule,

$$\delta w_{n+1} = -\eta \frac{\partial \mathcal{H}}{\partial w} + \alpha\, \delta w_n. \tag{17}$$

Assume that the system is initially at point A, and that $\eta s = 1/2$. The slope of the segment AB in Fig. 4 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?

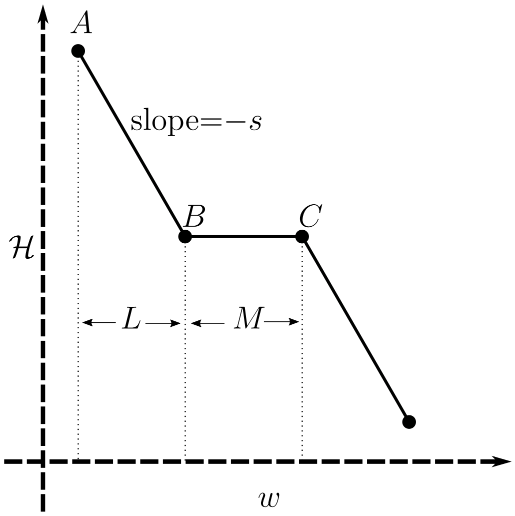

Figure 4: Energy as a function of weight for problem: Gradient descent and momentum.

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

$$\delta w_{i+1} = \sum_{j=0}^{i} \eta s\, \alpha^j + \alpha^{i+1} \delta w_0, \tag{18}$$

$$= \eta s \frac{1 - \alpha^{i+1}}{1 - \alpha}. \tag{19}$$

Next compute Δw_n,

$$\Delta w_n = \sum_{i=1}^{n} \delta w_i, \tag{20}$$

$$= \eta s \sum_{i=1}^{n} \frac{1-\alpha^{i+1}}{1-\alpha}, \tag{21}$$

$$= \frac{\eta s}{1-\alpha}\left(n - \alpha\frac{1-\alpha^n}{1-\alpha}\right). \tag{22}$$

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,

$$n_{\alpha=0} = 2L, \tag{23}$$

$$n_{\alpha=1/2} - 1 + 2^{-n_{\alpha=1/2}} = L. \tag{24}$$

Plotting these relations, we see that $n_{\alpha=1/2} < n_{\alpha=0}$, thus, $\alpha = 1/2$ converges faster.

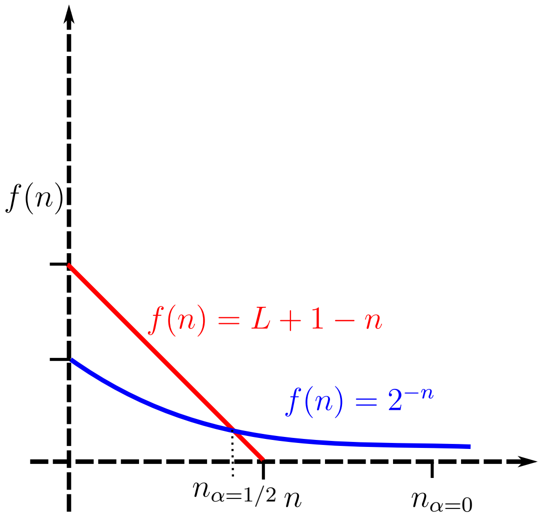

Figure 5: Graphical solution of problem : gradient descent and momentum.

After crossing point B, $\delta w(\alpha = 0) = 0$. The weights cease to change. On the other hand, $\delta w_{\alpha=1/2} > 0$. The weights keep changing.

6. Linear activation function Consider using a linear activation function $g(b) = b$ in a fully connected simple perceptron with one output unit. Fed with a training pattern $\boldsymbol{x}^{(\mu)}$, the output $O^{(\mu)}$ is given by

$$O^{(\mu)} = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}^{(\mu)} - \theta. \tag{25}$$

Here $\boldsymbol{w}$ is a column vector of weights, and θ is a scalar threshold. There are p training patterns, $\mu = 1, \ldots, p$. Their target outputs are denoted by $t^{(\mu)}$.

For the perceptron concidered, the energy function

$$H = \frac{1}{2} \sum_{\mu=1}^{p} \left(O^{(\mu)} - t^{(\mu)} \right)^2 \tag{26}$$

has only one minimum, and it can be found analytically. In the following, you will derive the threshold θ at the minimum.

a) Start by showing that the minimum implies

$$\mathbb{G}\boldsymbol{w} = \boldsymbol{\alpha} + \theta\boldsymbol{\beta} \tag{27a}$$

$$\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{w} = \theta + \gamma \tag{27b}$$

with

$$\mathbb{G} = \left\langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \right\rangle, \quad \boldsymbol{\alpha} = \langle t\boldsymbol{x} \rangle, \quad \boldsymbol{\beta} = \langle \boldsymbol{x} \rangle \quad \text{and} \quad \gamma = \langle t \rangle, \tag{28}$$

where $\langle \ldots \rangle$ denotes an average over the training patterns.

b) Assume that $\mathbb{G}$ is invertible, with inverse $\mathbb{G}^{-1}$. Furthermore, assume that $\boldsymbol{\beta}^{\mathsf{T}}\mathbb{G}^{-1}\boldsymbol{\beta} \neq 1$ and solve eqs. (27) for θ.

c) If, in a fully connected multi-layer perceptron, one uses a linear activation function $g(b) = b$, it holds that

$$\begin{aligned} \boldsymbol{V}^{(\mu,\ell)} &= \boldsymbol{w}^{(\ell)}\boldsymbol{V}^{(\mu,\ell-1)} - \boldsymbol{\theta}^{(\ell)} \\ &= \left[\boldsymbol{w}^{(\ell)}\boldsymbol{w}^{(\ell-1)}\right]\boldsymbol{V}^{(\mu,l-2)} - \left[\boldsymbol{w}^{(l)}\boldsymbol{\theta}^{(\ell-1)} + \boldsymbol{\theta}^{(\ell)}\right]. \end{aligned} \tag{29}$$

Here, $\boldsymbol{V}^{(\mu,\ell)}$ is the μ^{th} neuron in the ℓ^{th} hidden layer. Furthermore, $\boldsymbol{w}^{(\ell)}$ and $\boldsymbol{\theta}^{(\ell)}$ are the weight matrix and theshold vector for the neurons in the ℓ^{th} hidden layer. Write at most three sentences where you, based on eq. (29), argue that a non-linear activation function is essential for a multi-layer perceptron.

Solution: a)

$$
\begin{aligned}
\frac{\partial H}{\partial w_i} =& \frac{\partial}{\partial w_i}\frac{1}{2}\sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)^2 \\
=& \sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)\frac{\partial O^{(\mu)}}{\partial w_i} \\
=& \sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)x_i^{(\mu)} \\
=& \sum_{\mu=1}^{p}\left(\sum_{j=1}^{N} w_j x_j^{(\mu)} - \theta - t^{(\mu)}\right)x_i^{(\mu)} \\
=& \sum_{\mu=1}^{p}\left(\sum_{j=1}^{N} w_j x_j^{(\mu)}\right)x_i^{(\mu)} + \sum_{\mu=1}^{p}\left(-\theta\right)x_i^{(\mu)} + \sum_{\mu=1}^{p}\left(-t^{(\mu)}\right)x_i^{(\mu)} \\
=& \sum_{\mu=1}^{p}\sum_{j=1}^{N} w_j x_j^{(\mu)}x_i^{(\mu)} - \sum_{\mu=1}^{p}\theta x_i^{(\mu)} - \sum_{\mu=1}^{p}t^{(\mu)}x_i^{(\mu)} \\
=& \sum_{j=1}^{N}\sum_{\mu=1}^{p} w_j x_j^{(\mu)}x_i^{(\mu)} - \theta\sum_{\mu=1}^{p} x_i^{(\mu)} - \sum_{\mu=1}^{p}t^{(\mu)}x_i^{(\mu)} \\
=& \sum_{j=1}^{N} w_j \sum_{\mu=1}^{p} x_j^{(\mu)}x_i^{(\mu)} - \theta\sum_{\mu=1}^{p} x_i^{(\mu)} - \sum_{\mu=1}^{p}t^{(\mu)}x_i^{(\mu)} \\
=& \sum_{j=1}^{N} w_j p G_{ji} - \theta p \beta_i - p\alpha_i = p\left(\sum_{j=1}^{N} G_{ij}w_j - \theta\beta_i - \alpha_i\right)
\end{aligned}
$$

$$\frac{\partial H}{\partial w_i} = 0 \Rightarrow \mathbb{G}\boldsymbol{w} = \boldsymbol{\alpha} + \theta\boldsymbol{\beta} \tag{30}$$

$$\begin{aligned}
\frac{\partial H}{\partial \theta} =& \frac{\partial}{\partial \theta}\frac{1}{2}\sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)^2 \\
=& \sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)\frac{\partial O^{(\mu)}}{\partial \theta} \\
=& \sum_{\mu=1}^{p}\left(O^{(\mu)} - t^{(\mu)}\right)x_i^{(\mu)} \\
=& \sum_{\mu=1}^{p}\left(\sum_{j=1}^{N} w_j x_j^{(\mu)} - \theta - t^{(\mu)}\right)(-1) \\
=& -\sum_{\mu=1}^{p}\left(\sum_{j=1}^{N} w_j x_j^{(\mu)}\right) - \sum_{\mu=1}^{p}(-\theta) - \sum_{\mu=1}^{p}(-t^{(\mu)}) \\
=& -\sum_{\mu=1}^{p}\sum_{j=1}^{N} w_j x_j^{(\mu)} + \sum_{\mu=1}^{p}\theta + \sum_{\mu=1}^{p} t^{(\mu)} \\
=& -\sum_{j=1}^{N} w_j \sum_{\mu=1}^{p} x_j^{(\mu)} + p\theta + p\gamma \\
=& -p\sum_{j=1}^{N} w_j \beta_j + p\theta + pc
\end{aligned}$$

$$\frac{\partial H}{\partial \theta} = 0 \Rightarrow \boldsymbol{w}^{\mathsf{T}}\boldsymbol{\beta} = \theta + \gamma. \tag{31}$$

b) The first equation gives:

$$\boldsymbol{w} = \mathbb{G}^{-1}\boldsymbol{\alpha} + \theta\mathbb{G}^{-1}\boldsymbol{\beta}. \tag{32}$$

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

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

c) The equation can be written as

$$\boldsymbol{V}^{(\mu,\ell)} = \boldsymbol{W}\boldsymbol{V}^{(\mu,\ell-2)} - \boldsymbol{\Theta}, \tag{33}$$

where

$$\boldsymbol{W} = \boldsymbol{w}^{(\ell)}\boldsymbol{w}^{(\ell-1)}, \tag{34}$$

and

$$\boldsymbol{\Theta} = \boldsymbol{w}^{(l)}\boldsymbol{\theta}^{(\ell-1)} + \boldsymbol{\theta}^{(\ell)}. \tag{35}$$

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.