Solutions to exercises in Chapter 4

4.10 Thresholds in restricted Boltzmann machines. Consider first a Boltzmann machine without hidden neurons, but with thresholds. In this case, Equation (**4.16**) takes the form

$$P_{\mathrm{B}}(\boldsymbol{s}=\boldsymbol{x}) = Z^{-1}\exp\Big(\tfrac{1}{2}\sum_{i\neq j} w_{ij}x_i x_j - \sum_i \theta_i x_i\Big). \tag{1}$$

To derive the learning rule for the thresholds, we need to evaluate the gradient of

$$\frac{\partial \log \mathscr{L}}{\partial \theta_m} = \frac{\partial}{\partial \theta_m}\sum_\mu\Big(-\log Z + \tfrac{1}{2}\sum_{i\neq j} w_{ij}x_i^{(\mu)}x_j^{(\mu)} - \sum_i \theta_i x_i^{(\mu)}\Big), \tag{2}$$

with

$$\log Z = \sum_{s_!=\pm1,\ldots,s_N=\pm1} \exp\Big(\tfrac{1}{2}\sum_{i\neq j} w_{ij}s_i s_j - \sum_i \theta_i s_i\Big). \tag{3}$$

The derivative of $\log Z$ evaluates to

$$\frac{\partial \log Z}{\partial \theta_m} = -\sum_{s_!=\pm1,\ldots,s_N=\pm1} s_m P_{\mathrm{B}}(\boldsymbol{s}) = -\langle s_m\rangle_{\mathrm{model}} \tag{4}$$

Evaluating the derivative of the second term in Equation (2) in a similar way, one obtains

$$\frac{\partial \log \mathscr{L}}{\partial \theta_m} = -p\big(\langle x_m\rangle_{\mathrm{data}} - \langle s_m\rangle_{\mathrm{model}}\big). \tag{5}$$

Comparing with Equation (**4.26**), we see that the same rule of thumb applies as described in Chapter **6.1**: the learning rule for the thresholds is obtained from that of the weights by replacing the the state of the neuron in the weight-update formula by -1.

Now consider a restricted Boltzmann machine with hidden neurons. There are two thresholds in Equation (**4.29**), for the visible and for the hidden neurons.

$$\frac{\partial \mathscr{L}}{\partial \theta_n^{(\mathrm{v})}} = -(), \tag{6}$$

and

$$\frac{\partial \mathscr{L}}{\partial \theta_m^{(\mathrm{h})}} = -(). \tag{7}$$

4.11 Restricted Boltzmann machine with 0/1 neurons. We start from Equation (**4.32**),

$$\delta w_{mn}^{(\mu)} = \eta\big(\langle h_m x_n^{(\mu)}\rangle_{\text{data}} - \langle h_m v_n\rangle_{\text{model}}\big). \tag{8}$$

The term $\langle h_m x_n^{(\mu)}\rangle_{\text{data}}$ is computed by averaging over all states of the hidden neurons when the pattern $\boldsymbol{x}^{(\mu)}$ is clamped to the visible neurons. So

$$\langle h_m x_n^{(\mu)}\rangle_{\text{data}} = \sum_{h_1=0,1,\ldots,h_M=0,1} h_m x_n^{(\mu)} \Big[\prod_{i=1}^{M} P(h_i|\boldsymbol{v} = \boldsymbol{x}^{(\mu)})\Big]. \tag{9}$$

Using normalisation, $\sum_{h_j=0,1} P(h_j|\boldsymbol{v}=\boldsymbol{x}^{(\mu)}) = 1$, one finds

$$\langle h_m x_n^{(\mu)}\rangle_{\text{data}} = \sum_{h_m=0,1} h_m x_n^{(\mu)} P(h_m|\boldsymbol{v} = \boldsymbol{x}^{(\mu)}). \tag{10}$$

For 0/1 neurons, the stochastic update rule (**4.30**) is replaced by

$$h'_m = \begin{cases} 1 & \text{with probability} \quad p(b_m^{(\text{h})}), \\ 0 & \text{with probability} \quad 1 - p(b_m^{(\text{h})}), \end{cases} \tag{11}$$

with $b_m^{(\text{h})} = \sum_j w_{ij} v_j - \theta_i^{(\text{h})}$ and $p(b_m^{(\text{h})}) = [1 + \exp(-b_m^{(\text{h})})]^{-1}$. Note that the argument of the exponential functions lacks a factor of two, compared with Equation (**3.1**). We use (11) to evaluate the average in Equation (10):

$$\langle h_m x_n^{(\mu)}\rangle_{\text{data}} = p(b_m^{(\text{h})}). \tag{12}$$

The second average in (8) is evaluated in an analogous fashion

$$\langle h_m v_n\rangle_{\text{model}} = \langle p(b_m^{(\text{h})}) v_n\rangle_{\text{model}}. \tag{13}$$

Contrast Equations (12) and (13) with Equations (**4.34**) and (**4.35**). For ±1-neurons, the dependence b on the local field is $\tanh(b)$, just as in Equation (**3.7**). But for 0/1-neurons this is replaced by the sigmoid dependence $p(b)$.

Solutions to exercises in Chapter 6

6.2 Principal-component analysis. Consider first Figure 6.10. The patterns are

$$\boldsymbol{x}^{(1)}=\begin{bmatrix}-2\\-\frac{1}{2}\end{bmatrix},\quad \boldsymbol{x}^{(2)}=\begin{bmatrix}-1\\-\frac{1}{4}\end{bmatrix},\quad \boldsymbol{x}^{(3)}=\begin{bmatrix}1\\\frac{1}{4}\end{bmatrix},\quad \boldsymbol{x}^{(4)}=\begin{bmatrix}2\\\frac{1}{2}\end{bmatrix}. \tag{14}$$

Since the mean $\langle \boldsymbol{x}\rangle = p^{-1}\sum_{\mu=1}^{p}\boldsymbol{x}^{(\mu)}$ is zero, the elements of the covariance matrix are given by $C_{ij}=\langle x_i x_j\rangle$. We find

$$\mathbb{C}=\frac{1}{4}\begin{bmatrix}10 & \frac{5}{8}\\ \frac{5}{8} & \frac{5}{8}\end{bmatrix}. \tag{15}$$

The largest eigenvalue is $\lambda_1=85/32$, with eigenvector $\boldsymbol{u}_1\propto[4,1]^{\mathsf{T}}$. The second eigenvalue vanbishes, $\lambda_2=0$, because there is no data variance orthogonal to the principal direction.
The pattern vectors in Figure 6.11 are

$$\boldsymbol{x}^{(1)}=\begin{bmatrix}-6\\-5\end{bmatrix},\quad \boldsymbol{x}^{(2)}=\begin{bmatrix}-2\\-4\end{bmatrix},\quad \boldsymbol{x}^{(3)}=\begin{bmatrix}1\\3\end{bmatrix},\quad \boldsymbol{x}^{(4)}=\begin{bmatrix}2\\2\end{bmatrix},\quad \boldsymbol{x}^{(1)}=\begin{bmatrix}5\\4\end{bmatrix}, \tag{16}$$

Their mean vanishes, and the covariance matrix is given by

$$\mathbb{C}=\frac{1}{5}\begin{bmatrix}70 & 65\\ 65 & 70\end{bmatrix}. \tag{17}$$

Its largest eigenvalue is $\lambda_1=27$, and the corresponding eigenvector is $\boldsymbol{u}_1\propto[1,1]^{\mathsf{T}}$. This is the principal direction. The second eigenvalue is $\lambda_2=1$. It is not zero because the data in Figure 6.11 scatters a little bit around the principal direction.

6.5 Backpropagation. Consider first the learning rule for the output weights, W_{mn}. Using Equation (7.45), we find that the derivative of H w.r.t. W_{mn} evaluates to

$$\frac{\partial H}{\partial W_{mn}}=\sum_{i\mu}\frac{t_i^{(\mu)}-O_i^{(\mu)}}{O_i^{(\mu)}(1-O_i^{(\mu)})}\frac{\partial\sigma(B_i^{(\mu)})}{\partial W_{mn}}, \tag{18}$$

with $B_i^{(\mu)}=\sum_j W_{ij}V_j^{(\mu)}-\Theta_i$. We compute the derivative of σ using Equation (6.20). This gives

$$\frac{\partial\sigma(B_i^{(\mu)})}{W_{mn}}=\sigma(B_i^{(\mu)})[1-\sigma(B_i^{(\mu)})]\delta_{im}V_n^{(\mu)}. \tag{19}$$

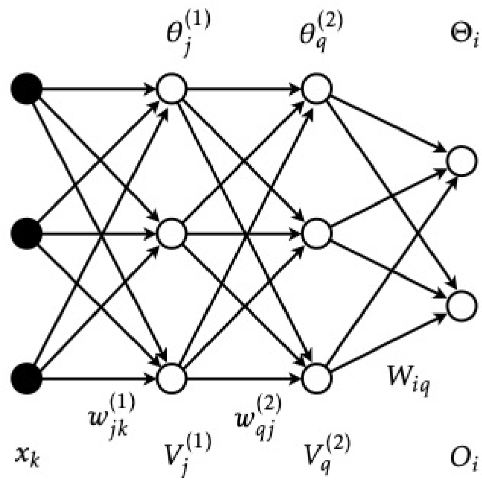

Figure 1: Network layout for Exercise 6.5. See also Figure 6.13.

This gives

$$\delta W_{mn} = -\eta \frac{\partial H}{\partial W_{mn}} = \eta \sum_\mu (t_m^{(\mu)} - O_m^{(\mu)}) V_n^{(\mu)} . \tag{20}$$

Now consider the learning rule for w_{mn}. Following the steps outlined in Section 6.1, one finds

$$\delta w_{mn} = \eta \sum_{i\mu} \Delta_i^{(\mu)} W_{im} \sigma'(b_m^{(\mu)}) x_n^{(\mu)} , \tag{21}$$

with $\Delta_i^{(\mu)} = t_i^{(\mu)} - O_i^{(\mu)}$. This expression differs from Equation(6.6b) by a factor of $\sigma'(B_i^{(\mu)})$.

6.6 Stochastic gradient descent. The network from Figure 6.13 is reproduced in Figure 1. How to derive the update formulae (or learning rules) for weights and thresholds is described in Section 6.1. The learning rules for the output weights and thresholds are simplest. For the weights, we have

$$\delta W_{mn} = -\eta \frac{\partial H}{\partial W_{mn}} = \eta \sum_\mu (t_m^{(\mu)} - O_m^{(\mu)}) g'(B_m^{(\mu)}) V_n^{(2,\mu)} , \tag{22}$$

corresponding to Equation (6.6a). The sequential learning rule is obtained by removing the sum over pattern indices μ. The learning rule for Θ_m is obtained from Equation (22) by setting $V_n^{(2,\mu)} = -1$ [Equation (6.11a)].
To find the learning rule for $w_{mn}^{(2)}$, we need to calculate

$$\frac{\partial O_i}{\partial w_{mn}^{(2)}} = \sum_q \frac{\partial O_i}{\partial V_q^{(2)}} \frac{\partial V_q^{(2)}}{\partial w_{mn}^{(2)}} . \tag{23}$$

Here we left out the pattern index μ. Using $\partial O_i/\partial V_q^{(2)} = g'(B_q)W_{iq}$ and $\partial V_q^{(2)}/\partial w_{mn}^{(2)} = g'(b_q^{(2)})\delta_{qm}V_n^{(1)}$, we find

$$\delta w_{mn}^{(2)} = \eta \sum_i (t_i - O_i) g'(B_m) W_{im} g'(b_m^{(2)}) V_m^{(1)} . \tag{24}$$

This is equivalent to Equation (6.8). The learning rule for $\theta_m^{(2)}$ is obtained upon replacing $V_m^{(1)}$ by -1.
The learning rule for $w_{mn}^{(1)}$ requires one more application of the chain rule

$$\frac{\partial O_i}{\partial w_{mn}^{(1)}} = \sum_q \frac{\partial O_i}{\partial V_q^{(2)}} \sum_j \frac{\partial V_q^{(2)}}{\partial V_j^{(1)}} \frac{\partial V_j^{(1)}}{\partial w_{mn}^{(1)}} . \tag{25}$$

Using $\partial V_q^{(2)}/\partial V_j^{(1)} = g'(b_q^{(2)}) w_{qj}^{(2)}$ and $\partial V_j^{(1)}/\partial w_{mn}^{(1)} = g'(b_j^{(1)})\delta_{jm} x_n$, we find

$$\delta w_{mn}^{(1)} = \eta \sum_i (t_i - O_i) \sum_q g'(B_q) W_{iq} g'(b_q^{(2)}) w_{qm}^{(2)} g'(b_m^{(1)}) x_n . \tag{26}$$

The learning rule for $\theta_m^{(1)}$ is obtained by setting $x_n = -1$.

6.8 Error backpropagation. To derive Equation (6.16), we start from

$$\delta w_{mn}^{(\ell)} = -\eta \frac{\partial H}{\partial w_{mn}^{(\ell)}} \quad \text{with} \quad H = \tfrac{1}{2} \sum_i \big(t_i - V_i^{(L)}\big)^2 . \tag{27}$$

Here we left out the sum over pattern indices μ in H, in order to get the stochastic gradient-descent algorithm (Sections 6.1 and 6.2). Evaluating the derivative yields

$$\delta w_{mn}^{(\ell)} = \eta \sum_i \big(t_i - V_i^{(L)}\big) \frac{\partial V_i^{(L)}}{\partial w_{mn}^{(\ell)}} = \eta \sum_i \big(t_i - V_i^{(L)}\big) \sum_q \frac{\partial V_i^{(L)}}{\partial V_q^{(\ell)}} \frac{\partial V_q^{(\ell)}}{\partial w_{mn}^{(\ell)}} , \tag{28}$$

where we applied the chain rule twice. Equation (6.14) allows us to compute the right-most derivative

$$\frac{\partial V_q^{(\ell)}}{\partial w_{mn}^{(\ell)}} = g'(b_q^{(\ell)}) \delta_{mq} V_n^{(\ell-1)} . \tag{29}$$

In summary,

$$\delta w_{mn}^{(\ell)} = \eta \sum_i \big(t_i - V_i^{(L)}\big) g'(b_m^{(\ell)}) \frac{\partial V_i^{(L)}}{\partial V_m^{(\ell)}} V_n^{(\ell-1)} . \tag{30}$$

Comparing with Equation (**6.15**), $\delta w_{mn}^{(\ell)} = \eta \delta_m^{(\ell)} V_n^{(\ell-1)}$, we find

$$\delta_m^{(\ell)} = \sum_i \big(t_i - V_i^{(L)}\big) \frac{\partial V_i^{(L)}}{\partial V_m^{(\ell)}} g'(b_m^{(\ell)}). \tag{31}$$

This is equivalent to Equation (**6.16**),

$$\delta_j^{(\ell-1)} = \sum_i \big(t_i - V_i^{(L)}\big) \frac{\partial V_i^{(L)}}{\partial V_j^{(\ell-1)}} g'(b_j^{(\ell-1)}), \tag{32}$$

which answers the first part of the question. To derive the recursion (**6.17**), we use the chain rule once more,

$$\frac{\partial V_i^{(L)}}{\partial V_j^{(\ell-1)}} = \sum_q \frac{\partial V_i^{(L)}}{\partial V_q^{(\ell)}} \frac{\partial V_q^{(\ell)}}{\partial V_j^{(\ell-1)}} = \sum_q \frac{\partial V_i^{(L)}}{\partial V_q^{(\ell)}} g'(b_q^{(\ell)}) w_{qj}^{(\ell)}. \tag{33}$$

Substituting this expression into Equation (32) gives

$$\begin{aligned}
\delta_j^{(\ell-1)} &= \sum_i \big(t_i - V_i^{(L)}\big) \sum_q \frac{\partial V_i^{(L)}}{\partial V_q^{(\ell)}} g'(b_q^{(\ell)}) w_{qj}^{(\ell)} g'(b_j^{(\ell-1)}), \qquad (34)\\
&= \sum_q \left(\sum_i \big(t_i - V_i^{(L)}\big) \frac{\partial V_i^{(L)}}{\partial V_q^{(\ell)}} g'(b_q^{(\ell)}) \right) w_{qj}^{(\ell)} g'(b_j^{(\ell-1)}).
\end{aligned}$$

The last step is to compare with Equation (31). This yields Equation (**6.17**):

$$\delta_j^{(\ell-1)} = \sum_q \delta_q^{(\ell)} w_{qj}^{(\ell)} g'(b_j^{(\ell-1)}). \tag{35}$$