Problem 1: Backpropagation.

$$\delta W_{mn} = -\eta \frac{\partial H}{\partial W_{mn}} = -\eta \sum_{p} \frac{\partial H}{\partial O_p} \frac{\partial O_p}{\partial W_{mn}}$$

$$= -\eta \sum_p \frac{\partial H}{\partial O_p} \sigma'(B_p) \sum_q \frac{\partial W_{pq}}{\partial W_{mn}} X_q$$

$$= +\eta \sum_{\mu,p} \left(\frac{t_p^{(\mu)}}{O_p^{(\mu)}} - \frac{(1-t_p^{(\mu)})}{1-O_p^{(\mu)}} \right) \sigma'(B_p) \sum_q \delta_{pm}\delta_{qn} X_q$$

$$= \eta \sum_\mu \frac{\left(t_m^{(\mu)} - O_m^{(\mu)}\right)}{O_m^{(\mu)}\left(1-O_m^{(\mu)}\right)} \sigma(B_m)(1-\sigma(B_m)) X_n$$

using $O_m^{(\mu)} = \sigma(B_m^{(\mu)})$

$$= \eta \sum_\mu \left(t_m^{(\mu)} - O_m^{(\mu)}\right) X_n$$

Similarly by chain rule, obtain.

$$\delta w_{mn} = \eta \sum_{\mu,p} \left(t_p^{(\mu)} - O_p^{(\mu)}\right) W_{pm} \sigma'(b_m) x_n$$

Problem 2: Restricted Boltzmann Machine.

Definitions: η : learning rate

$\delta w_{mn}^{(\mu)}$: weight update for the weight w_{mn}, corresponding to pattern "μ".

h_m : state of the m^{th} hidden neuron

$x_n^{(\mu)}$: input

v_n : state of the n^{th} visible neuron.

How to perform the averages:

$\langle h_m x_n^{(\mu)} \rangle_{data}$ is computed by averaging over all states of the hidden neurons, given that pattern "μ" is applied to the visible neurons.

$\langle h_m v_n \rangle_{model}$ can be simplified in a similar manner to the $\langle \cdot \rangle_{data}$ average, but is computed in simulations by M.C. sampling.

$$\langle h_m x_n^{(\mu)} \rangle_{data} = \sum_{h_i=0,1} h_m x_n^{(\mu)} \prod_{i=1} P(h_i | v = x^{(\mu)})$$

$$= \sum_{h_m=0,1} h_m P(h_m | v = x)$$

$$= 1 \cdot P(b_m^{(\mu)}) + 0\,(1 - P(b_m^{(\mu)}))$$

$$= P(b_m^{(\mu)})$$

where $P(b_m^{(\mu)}) = \dfrac{1}{1 + e^{-2 b_m^{(\mu)}}}$

Similarly,

$$\langle h_m v_n \rangle_{model} = \left\langle \frac{v_n}{1 + e^{-2 b_m^{(\mu)}}} \right\rangle_{model}.$$

Contrast: we find a sigmoid dependence on $b_m^{(\mu)}$ instead of a tanh dependence, reflecting that the 0,1 neurons do not have a "0" mean. Whereas 1/-1 neurons do.

2.

Choose the following filters:

1. $\begin{bmatrix} +1 & +1 \\ -1 & -1 \end{bmatrix}$ 2. $\begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$ 3. $\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$

4. $\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$

Choose thresholds = 0, stride = 1, padding = 0.

Use max pooling, 2x2. layers

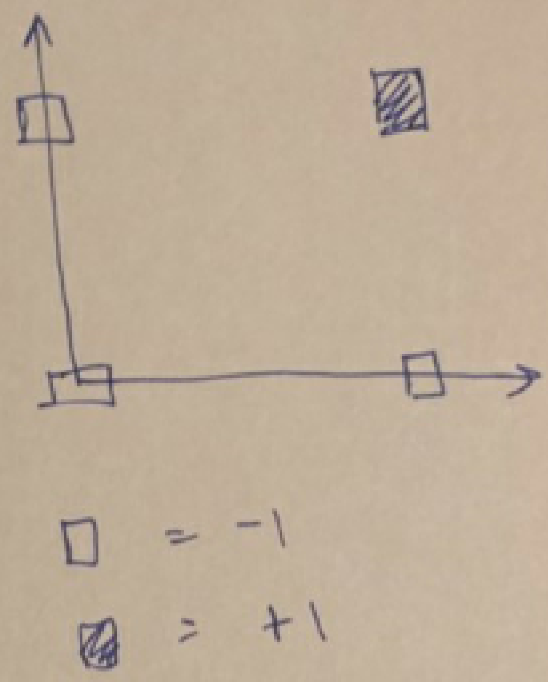
Let bars be monochromatic columns, and stripes be monochromatic rows.

for bars, filters 1, 2 (+ max pool) output 0, whereas 3, 4 + max pool output 2, and vice versa for stripes. Let x_i be the output of the i^{th} filter + max pool.

Then a F.C. layer, $g(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 - \theta)$ with $w_1 = w_2 = +1$, $w_3 = w_4 = -1$, $\theta = 0$, and step activation "g" outputs 0 for bars & 1 for stripes.

4. a) Boolean AND problem

Input		Output
0	0	-1
0	1	-1
1	0	-1
1	1	1



b) $$H = \frac{1}{2}\Big[(1 - w_1 - w_2 + \theta)^2 + (-1 - w_1 + \theta)^2 + (-1 - w_2 + \theta)^2 + (-1 + \theta)^2\Big]$$

Set $\frac{\partial H}{\partial w_1}, \frac{\partial H}{\partial w_2}, \frac{\partial H}{\partial \theta} = 0$.

$$\frac{\partial H}{\partial w_1} = 2w_1 + w_2 - 2\theta = 0$$

$$\frac{\partial H}{\partial w_2} = w_1 + 2w_2 - 2\theta = 0$$

$$\frac{\partial H}{\partial \theta} = -2 - 2w_1 - 2w_2 + 4\theta = 0$$

Solution: $w_1 = 1$, $w_2 = 1$, $\theta = 3/2$

(c) See Book section 5.3.
reason: overlap matrix not invertible for l.d. inputs.

(d) No, do not solve the Boolean AND.
The above procedure is equivalent to computing the pseudo inverse for the overlap matrix.
Thus essentially works as a best fit

(5). Autoencoders

a. write output of a layer as $b_i^{(\mu)} = w_{ij} x_j^{(\mu)} - \theta_i$
then, average over μ ~~and~~ ~~set~~ ~~assume~~ inputs ~~have mean~~,

$$\langle b_i^{(\mu)} \rangle_\mu = w_{ij} \cancelto{0}{\langle x_j^{(\mu)} \rangle_\mu} - \theta_i$$

$$\langle b_i^{(\mu)} \rangle_\mu = -\theta_i$$

choosing $\theta_i = 0$, sets $\langle b_i^{(\mu)} \rangle_\mu = 0$

b) $$H = \frac{1}{2} \sum_{i,\mu} \left(t_i^{(\mu)} - \sum_{j,k} (W_d)_{ij} (W_e)_{jk} x_k^{(\mu)} \right)^2$$

$$E = \frac{1}{2} \| X - W_d W_e X \|^2$$

Write $Y_h = W_e X$

$X = n \times N$

assume that X has rank n.

In the given case, $W_d = n \times 2$, $n > 2$,
so can have rank at most 2.

the matrix $W_d Y_h$ that minimises E
is the ~~is~~ best rank 2 approximation of X.
can be solved using SVD.

$$X = U_n \Sigma_n V_n^T$$

$$W_d Y_h = U_p \Sigma_p V_p \; ; \quad p = 2.$$

thus one ~~the~~ solution can be written

$$W_d = U_p T^{-1} \; ; \; Y_h = T \Sigma_p V_p$$

for arbit. $p \times p$ T.

finally, need to solve

$$\underset{p\times n}{W_e}\ \underset{n\times N}{X} = \underset{p\times p}{\Pi}\ \underset{p\times p}{\Sigma_p}\ \underset{p\times N}{V_p^T}$$

takes SVD pseudo inverse of X

$$W_e = \underset{p\times p}{\Pi}\ \underbrace{\underset{p\times N}{\Sigma_p}\ \underset{N\times n}{V_p^T}\ \underset{n\times n}{V_n}\ \underset{n\times n}{\Sigma_n^{-1}}\ U_n^T}_{\mathbb{1}_{p\times n}}$$

6. $P^{t>1}_{error} = \text{Prob}\left(C_i^{(\nu)} > 1\right)$

Using CLT, $C_i^{(\nu)}$ is normally distributed with mean $= 0$ & variance $\sigma_c^2 \approx \frac{P}{N}$

$$P^{t>1}_{error} = \int_1^\infty dc \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{c^2}{2\sigma_c^2}}$$

$= \frac{1}{2}\left[1 - \text{erf}\left(\sqrt{\frac{1}{2\sigma_c^2}}\right)\right]$