CHALMERS, GÖTEBORGS UNIVERSITET

EXAM for
ARTIFICIAL NEURAL NETWORKS

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

Maximum score on this exam: 12 points. Each question gives at most 2 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 >13.5 passed; >17 grade 4; >21.5 grade 5,
GU >13.5 grade G; > 19.5 grade VG.

---

1. Backpropagation
Derive stochastic gradient-descent learning rules for the weights of the network shown in Figure 1. All activation functions are of sigmoid form, $\sigma(b) = 1/(1+\mathrm{e}^{-b})$, hidden thresholds are denoted by θ_j, and those of the output neurons by Θ_i. The energy function is

$$H = -\sum_{i,\mu} \left[t_i^{(\mu)} \log O_i^{(\mu)} + (1 - t_i^{(\mu)}) \log(1 - O_i^{(\mu)}) \right] , \qquad (1)$$

where log is the natural logarithm, $t_i^{(\mu)}$ are the targets, $O_i^{(\mu)}$ are the outputs, and μ labels different inputs.

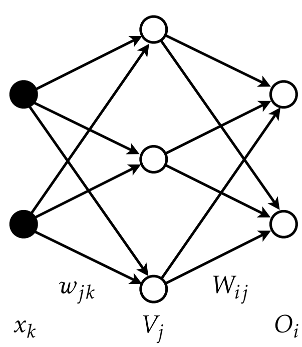

Figure 1: Network layout for question 1.

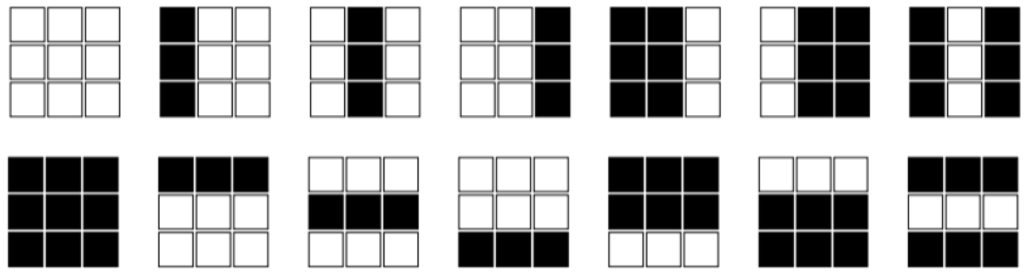

Figure 2: Bars-and-stripes ensemble, ■ corresponds to $x = 1$, and □ to $x = 0$. Question 3

2. Restricted Boltzmann machine with 0/1 neurons
The learning rule for a restricted Boltzmann machine reads:

$$\delta w_{mn}^{(\mu)} = \eta \big(\langle h_m x_n^{(\mu)} \rangle_{\text{data}} - \langle h_m v_n \rangle_{\text{model}} \big) \,. \qquad (2)$$

Give the definitions of all variables in Equation (2), explain how the averages $\langle \cdots \rangle_{\text{data}}$ and $\langle \cdots \rangle_{\text{model}}$ are computed, and evaluate them further by performing the sums over the states of the hidden neurons contained in the averages. Assume that the states of all neurons take the values 0 or 1. Contrast your results with those for ± 1 neurons, given e.g. in the lecture notes.

3. Convolutional net
Figure 2 shows the patterns of the 3×3 bars-and-stripes ensemble with 0/1 digits. Construct a convolutional network that classifies the patterns into bars and stripes. Use one convolution layer with at most four 2×2 kernels, one pooling layer, and at most one fully connected layer for classification. Give all parameters of the network (all weights, thresholds, activation functions, padding, and stride where relevant).

4. Boolean AND problem
(**a**) Consider a linear unit with weight vector $\boldsymbol{w}$ and threshold θ. Choose appropriate inputs $\boldsymbol{x}^{(\mu)}$ and targets $t^{(\mu)}$ to represent the Boolean AND function. Write down the corresponding value table, and draw the patterns in the input plane.

(**b**) Set the gradients of the energy function $H = \frac{1}{2}\sum_\mu (t^{(\mu)} - O^{(\mu)})^2$ to zero and solve the resulting equations for the weights and the thresholds.

(**c**) Explain why a linear unit cannot solve binary classification problems when the inputs are linearly dependent.

(**d**) Do the weights and thresholds you obtained in (**b**) solve the Boolean AND problem? After all, those values correspond to an extremum of the energy function. Discuss (only a couple of sentences needed).

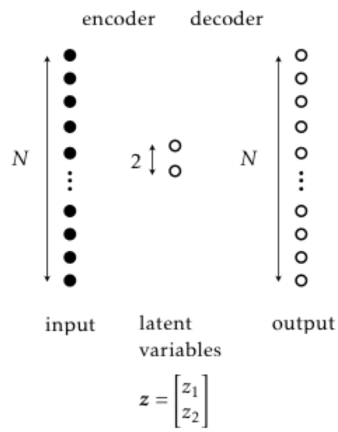

Figure 3: Autoencoder, question 5. Input and output dimension N.

5. Autoencoder
Figure 3 shows the layout for an autoencoder. The bottleneck layer has two neurons, their states z_1 and z_2 are the latent variables. All neurons are linear units with zero thresholds.

(**a**) Explain why it is ok to put the thresholds to zero for zero-mean inputs.

(**b**) Write down the quadratic energy function in terms of the weight matrices $\mathbb{W}_{\rm e}$ and $\mathbb{W}_{\rm d}$ of the encoder and the decoder.

(**c**) Derive the optimal weight matrix of the decoder. Use singular-value decomposition of the matrix $\mathbb{X}$ that has the pattern vectors $\boldsymbol{x}^{(\mu)}$ as its columns.

(**d**) Derive the optimal weight matrix for the encoder. Analyse the resulting expression for $\boldsymbol{z}^{(\mu)}$ and argue that the latent variables $z_1^{(\mu)}$ and $z_2^{(\mu)}$ are the top two principal components of the input $\boldsymbol{x}^{(\mu)}$ (assuming zero-mean input data and optimal weights).

6. Error probability in Hopfield model
(**a**) Derive the expression

$$P_{\rm error}^{t=1} = \frac{1}{2}\Big[1 - {\rm erf}\Big(\sqrt{\frac{1}{2\alpha}}\Big)\Big] \tag{3}$$

for the one-step error probability for the deterministic Hopfield model, where α is the storage capacity. Start by feeding a stored pattern, and compute the probability that a bit is erroneously changed under the McCulloch-Pitts dynamics. Use the central-limit theorem (do **not** derive this theorem, but state what it says, and how it is used in your derivation).

(**b**) The cross-talk term is negligible if the storage capacity is small enough. The condition is $\alpha \ll N^{-1}\log N$, where N is the number of bits per pattern. Where does the factor $\log N$ come from (log is the natural logarithm)?