*Allowed equipment: pens, pencils, erasers, a non-programmable calculator*

Question 1 [max. 9 points]: Explain the following concepts briefly but informatively (each with 30–50 words, with a mathematical definition and/or with an illustration):

(i) activation function

(ii) bias-variance tradeoff

(iii) hyperparameters

(iv) learning rate

(v) recurrent neural networks

(vi) stochastic gradient descent

(vii) stochastic hidden unit

(viii) universal approximation theory

(ix) variational autoencoder
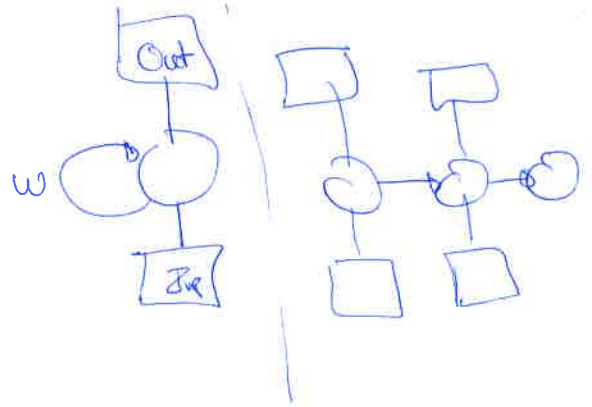
Question 2 [max. 3 points]:

Explain briefly why the selection of initial parameter values of a deep neural network is important. What are the common strategies for the parameter initialization?

Question 3 [max. 6 points]:

Explain verbally the principle and the motivation of the following three regularization methods. Also, enhance your explanation with mathematical expressions and/or graphical illustrations.

(i) weight decay

(ii) dropout

(iii) early stopping

Question 4 [max. 6 points]:

Assume a multilayer perceptron with two input layer units x_1 and x_2, one hidden layer with two units h_1 and h_2, and an output layer with one unit $\hat{y}$, with the activation (and the parameters) of each unit being scalar-valued.

Furthermore, let $w_{i,j}^{(1)}$ denote the connection weight between the units x_i and h_j, $w_j^{(2)}$ the connection weight between the units h_j and $\hat{y}$, $b_j^{(1)}$ the bias of the unit h_j, and $b^{(2)}$ the bias of the unit $\hat{y}$. The activation function in the hidden units is the logistic sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}} \, ,$$

and the activation function in the output unit is the linear function.

A graphical illustration of the very simple feedforward neural network:
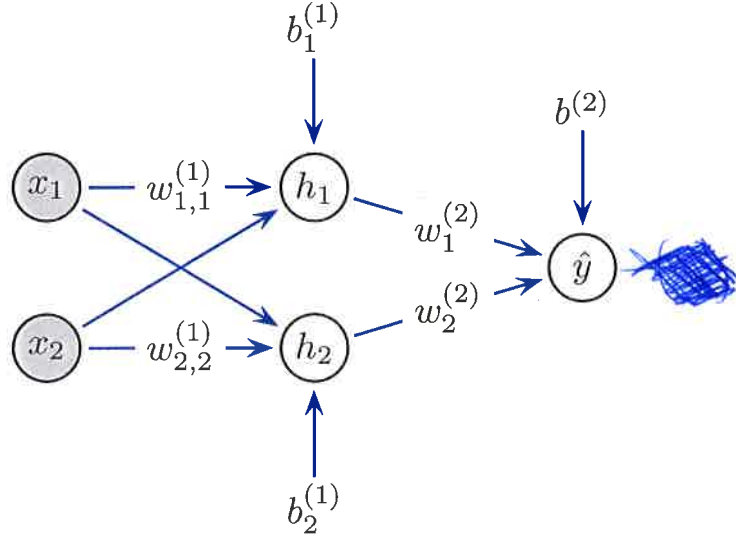
Finally, the training cost function applied to the network output $\hat{y}$ assuming target y is

$$J\left(x_1, x_2, y; \{w_{i,j}^{(1)}\}_{i,j=1}^2, \{b_j^{(1)}\}_{j=1}^2, \{w_j^{(2)}\}_{j=1}^2, b^{(2)}\right) = \frac{1}{2}(\hat{y} - y)^2 .$$

Derive the analytical expressions of the following:

(i) activations of the units h_1, h_2, and $\hat{y}$ (in forward propagation)

(ii) partial derivatives of the cost function with respect to the parameters $w_2^{(2)}$ and $w_{2,2}^{(1)}$, i.e., $\frac{\partial J}{\partial w_2^{(2)}}$ and $\frac{\partial J}{\partial w_{2,2}^{(1)}}$, using back-propagation

(iii) parameter update equations for the parameters $w_2^{(2)}$ and $w_{2,2}^{(1)}$, assuming a gradient descent update

Question 5 [max. 4 points]:

Name the typical building blocks of convolutional neural networks. Describe their functionality and motivation.

Question 6 [max. 5 points]:

Assume a binary (Bernoulli-Bernoulli) restricted Boltzmann machine having the following energy function:

$$E(\mathbf{v}, \mathbf{h}; \Theta) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_j \sum_i W_{j,i} h_j v_i ,$$

where $\mathbf{v} = \{v_i\}_{i=1}^I$ denotes the set of visible (observed) variable units, $\mathbf{h} = \{h_j\}_{j=1}^J$ the set of latent (hidden) variable units, and $\Theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ the set of the model's parameters. The free-energy function of the model is

$$F(\mathbf{v}; \Theta) = -\sum_i a_i v_i - \sum_j \log\left\{1 + \exp\left\{b_j + \sum_i W_{j,i} v_i\right\}\right\} .$$

Derive the analytical form of the probability distribution of $v_i | \mathbf{h}$, where $i \in \{1, \ldots, I\}$.