# CS551: Introduction to Deep Learning

**End Semester, Spring 2017**

**IIT Patna**

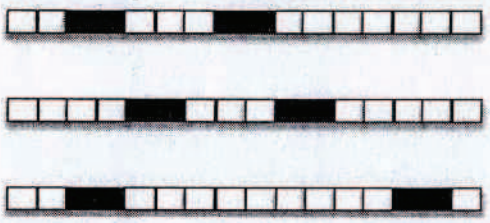
Attempt all questions. Do not write anything on the question paper.

Time: 3 Hrs Full marks: 50

1. Given a set of training example $(\boldsymbol{X}, \boldsymbol{y})$ where $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ are all the input examples and $\boldsymbol{y} \in \mathbb{R}^{m \times 1}$ be the corresponding output. We would like to fit a curve $(y = \boldsymbol{w}^T \boldsymbol{x})$ to predict $y \in \mathbb{R}$ for a given $\boldsymbol{x} \in \mathbb{R}^n$. Derive an expression, involving $\boldsymbol{X}$ and $\boldsymbol{y}$ only, to determine $\boldsymbol{w}$. (5)

2. What is early stopping? Show that it acts as regularizer. (3+5)

3. Consider the following image. We would like a develop a binary threshold discriminator to identify the existence of the pattern in the given input. The pattern can be translated or wrapped around at the boundary. Hence the pattern exists for all three following inputs. Is it possible to design a binary threshold discriminator to identify the pattern? If so, design it, otherwise provide necessary justifications. (3)



4. Given 3 data points in 2-d space, (1, 1), (2, 2) and (3, 3), (a) what is the first principle component? (b) If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data? (c) For the projected data in (b), now if we represent them in the original 2-d space, what is the reconstruction error? (2+1+1)

5. (a) What is gated recurrent unit (GRU)? What are the advantages and disadvantages of GRU with respect to long short term memory? (b) Describe very briefly back-propagation through time. (2+4)

6. Consider a neural network architecture consists of two layers - first layer uses convolution filters and the second layer does max-pooling. The input to the first layer is an image of size $227 \times 227 \times 3$. In the convolution layer there are 96 kernels each of size $11 \times 11 \times 3$ with stride of 4 and without 0 padding. (a) what will be the output size of first layer and what is the total number of parameters? (b) In the max-pool layer $3 \times 3 \times 1$ filter is applied. What will be the volume of output size and what is the total number of parameters in this layer? (2+2)

7. What is momentum? Describe briefly the steps for stochastic gradient descent with momentum. (3+3)

8. You want to train a neural network to drive a car. Your training data consists of gray scale $64 \times 64$ pixel images. The training labels include the human driver's steering wheel angle in degrees and the human driver's speed in miles per hour. Your neural network consists of an input layer with $64 \times 64 = 4,096$ units, a hidden layer with 2,048 units, and an output layer with 2 units (one for steering angle, one for speed). You use the

ReLU activation function for the hidden units and no activation function for the outputs (or inputs). (a) Calculate the number of parameters (weights) in this network. You can leave your answer as an expression. Be sure to account for the bias terms. (b) You train your network with the cost function $J = |\boldsymbol{y} - \boldsymbol{z}|^2$ . Use the following notation to derive $\frac{\partial J}{\partial W_{ij}}$.

- $\boldsymbol{x}$ is a training image (input) vector with a 1 component appended to the end, $y$ is a training label (input) vector, and $z$ is the output vector. All vectors are column vectors.
- $r(\gamma) = \max 0, \gamma$ is the ReLU activation function, $r'(\gamma)$ is its derivative (1 if $\gamma > 0$, 0 otherwise), and $r(\boldsymbol{v})$ is $r(\cdot)$ applied component-wise to a vector.
- $\boldsymbol{g}$ is the vector of hidden unit values before the ReLU activation functions are applied, and $\boldsymbol{h} = r(\boldsymbol{g})$ is the vector of hidden unit values after they are applied (but we append a 1 component to the end of $\boldsymbol{h}$).
- $\boldsymbol{V}$ is the weight matrix mapping the input layer to the hidden layer; $\boldsymbol{g} = \boldsymbol{V}\boldsymbol{x}$.
- $\boldsymbol{W}$ is the weight matrix mapping the hidden layer to the output layer; $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{h}$.

(c) Write $\frac{\partial J}{\partial W}$ as an outer product of two vectors. $\frac{\partial J}{\partial W}$ is a matrix with the same dimensions as $\boldsymbol{W}$; (d) Derive $\frac{\partial J}{\partial V_{ij}}$. (1+2+2+3)

9. Consider an Markovian Decision Process with states $4, 3, 2, 1, 0$, where 4 is the starting state. In states $k \geq 1$, you can walk (W) and $T(k, W, k-1) = 1$. In states $k \geq 2$, you can also jump (J) and $T(k, J, k-2) = T(k, J, k) = 1/2$. State 0 is a terminal state. The reward $R(s, a, s_0) = (s - s_0)^2$ for all $(s, a, s_0)$. Use a discount of $\gamma = 1/2$. (a) Compute $V^*(2)$. (b) Compute $Q^*(4, W)$. (3+3)