# DEEP LEARNING – WORKSHEET 5

## Q1 to Q8 are MCQs with only one correct answer. Choose the correct option.

1. Which of the following are advantages of batch normalization?
   1. Reduces internal covariant shift.
   2. Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
   3. allows use of saturating nonlinearities and higher learning rates.
   4. All of the above

**Answer:** D) All of the above

1. Which of the following is not a problem with sigmoid activation function?
   1. Sigmoids do not saturate and hence have faster convergence
   2. Sigmoids have slow convergence.
   3. Sigmoids saturate and kill gradients.
   4. Sigmoids are not zero centered; gradient updates go too far in different directions, making optimization more difficult.

**Answer:** A) Sigmoids do not saturate and hence have faster convergence

1. Which of the following is not an activation function?
   1. Swish B) Maxout

C) SoftPlus D) None of the above

**Answer:** A) Swish

1. The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False?
   1. True B) False

**Answer:** A) True

1. In which of the weights initialisation techniques, does the variance remain same with each passing layer?
   1. Bias initialisation B) Xavier Initialisation

C) He Normal Initialisation D) None of these

**Answer:** B) Xavier Initialisation

1. Which of the following is main weakness of AdaGrad?
   1. learning rate shrinks and becomes infinitesimally small
   2. learning rate doesn’t shrink beyond a point
   3. change in learning rate is not adaptive
   4. AdaGrad adapts updates to each individual parameter

**Answer:** A) learning rate shrinks and becomes infinitesimally small

1. In order to achieve right convergence faster, which of the following criteria is most suitable?
   1. momentum and learning rate both must be high
   2. momentum must be high and learning rate must be low
   3. momentum and learning rate both must be low
   4. momentum must be low and learning rate must be high

**Answer:** A) momentum and learning rate both must be high

1. When is an error landscape is said to be poor (ill) conditioned?
   1. when it has many local minima
   2. when it has many local maxima
   3. when it has many saddle points and flat areas
   4. None of these

**Answer:** A) when it has many local minima

## Q9 and Q10 are MCQs with one or more correct answers. Choose all the correct options.

1. Which of the following Gradient Descent algorithms are adaptive?
   1. ADAM B) SGD

C) NADAM D) RMS Prop

**Answer:** A) ADAM, C) NADAM, D) RMS Prop

1. When should an optimization function (gradient descent algorithm) stop training:
   1. when it reaches local minimum B) when it reaches saddle point
2. when it reaches global minimum
3. when it reaches a local minima which is similar to global minima (i.e. which has very less error distance with global minima)

**Answer:** C) When it reaches global minimum

## Q11 to Q15 are subjective answer type question. Answer them briefly.

1. What are convex, non-convex optimization?

**Answer:** A **convex optimization** problem is a problem where all of the constraints are convex functions, and the objective is a convex function if minimizing, or a concave function if maximizing. Linear functions are convex, so linear programming problems are convex problems.

A **non-convex optimization** may have multiple locally optimal points and it can take a lot of time to identify whether the problem has no solution or if the solution is global.

1. What do you mean by saddle point? Answer briefly.

**Answer:** Consider the function f(x,y)= - . Let's make a few observations about what goes on around the origin (0, 0).

* Both partial derivatives are 0 at this point.

- ) = 2x 🡪 2(0) =0

- ) = -2y 🡪 2(0) =0

Therefore (0,0) is a stable point.

* When you just move in the x direction around this point, the function looks like f(x,0)= – = . The single-variable function f(x)= has local minimum at x=0.
* When you just move in the y direction around this point, meaning the function looks like f(0,y)= – = . The single-variable function f(y)= - has local maximum at y=0.

In other words, the x and y directions disagree over whether this input should be a maximum or a minimum point. So even though (0, 0) is a stable point, and is not an inflection point, it cannot be a local maximum or local minimum

1. What is the main difference between classical momentum and Nesterov momentum? Explain briefly.

**Answer:** The main difference is in classical momentum you first correct your velocity and then make a big step according to that velocity (and then repeat), but in Nesterov momentum you first making a step into velocity direction and then make a correction to a velocity vector based on new location (then repeat).

1. What is Pre initialisation of weights? Explain briefly.

**Answer:** The aim of weight initialization is to prevent layer activation outputs from exploding or vanishing during the course of a forward pass through a deep neural network.

There are numerous weight initialization methods:

* One of the ways is to initialize all weights to 0s. As all the weights are same, the activations in all hidden units are also the same. This makes the gradient with respect to each weight be same.
* The other way is to initialize weights randomly from a uniform distribution. Every number in uniform distribution has equal probability to be picked.

1. What is internal covariance shift in Neural Networks?

**Answer:** Each layer takes some input, transforms this input through interaction with its weights, and outputs the result, to be consumed by the first layer downstream. The first layer will transform this data into something else. Statistically, however, this is also a sample, which thus has a sample mean and a sample standard deviation. This process repeats itself for each individual layer: the input data can be represented as some statistical sample with mean μ and standard deviation σ. Internal Covariate Shift as the change in the distribution of network activations due to the change in network parameters during training.