# 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:** D) None of the above

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:** B) momentum must be high and learning rate must be low

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:** C) when it has many saddle points and flat areas

## 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, D) When it reaches a local minima which is similar to global minima (i.e. which has very less error distance with global minima)

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

1. What are convex, non-convex optimization?

**Answer:** **Convex Optimization:**

It is an optimizing technique which involves a function in which there is only one optimum, corresponding to the global optimum (maximum or minimum). There is no concept of local optima for convex optimization problems, making them relatively easy to solve.

**Non-Convex Optimization:**

It is an optimization technique which involves a function which has multiple optima, only one of which is the global optima. Depending on the loss surface, it can be very difficult to locate the global optima.

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

**Answer:** Saddle points are the stable points which have local minima in one direction but local maxima in other direction.



In the adjacent figure, the point shown in black is a saddle point, It has a local minimum along the direction AB but has a local Maximum along the direction CD. Since, it looks like a saddle which is put over a horse, it is name d as a saddle point.

This is a non-convex function with a global minimum located within a long and narrow valley. Finding the valley is relatively easy, but it is difficult to converge to the global minimum due to the flat valley, which thus has small gradients so it is difficult for gradient-based optimization procedures to converge.

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). Simple we can say that in Nesterov momentum, the gradient step is adaptive.



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

**Answer:** Pre-Initialisation is common for convolutional networks used for examining images. The technique involves importing the weights of an already trained network (such as VGG16) and using these as the initial weights of the network to be trained.

This technique is only really viable for networks which are to be used on similar data to that which the network was trained on. For example, VGG16 was developed for image analysis, if you are planning to analyse images but have few data samples in your data set, pre-initialization might be a tenable method to utilize. This is the underlying concept behind transfer learning, but the terms pre-initialization and transfer learning are not necessarily synonymous.

1. What is internal covariance shift in Neural Networks?

**Answer:** In neural networks, the output of the first layer feeds into the second layer, the output of the second layer feeds into the third, and so on. When the parameters of a layer change, the distribution of inputs to subsequent layers also changes. We define Internal Covariate Shift as the change in the distribution of network activations due to the change in network parameters during training.

These shifts in input distributions can be problematic for neural networks, as it has a tendency to slow down learning, especially deep neural networks that could have a large number of layers.

It is well established that networks converge faster if the inputs have been whitened (ie zero mean, unit variances) and are uncorrelated and internal covariate shift leads to just the opposite.

**Batch normalization** is a method intended to mitigate internal covariate shift for neural networks.