# 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

Answers: ( A )

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: ( D )

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

C) SoftPlus D) None of the above

Answer: D )

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 )

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

C) He Normal Initialisation D) None of these

Answer: ( C )

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 )

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: ( C )

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 )

## 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.

Answe: (B) (D)

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:( A) (D)

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

1. What are convex, non-convex optimization?

Answer: **Convex optimization** is a subfield of mathematical **optimization** that studies the problem of minimizing **convex** functions over **convex** sets. Many classes of **convex optimization** problems admit polynomial-time algorithms, whereas mathematical **optimization** is in general NP-hard.

A **non**-**convex optimization** problem is any problem where the objective or any of the constraints are **non**-**convex**, as pictured below. Such a problem may have multiple feasible regions and multiple locally optimal points within each region.

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

Answer: a **point** at which a function of two variables has partial derivatives equal to zero but at which the function has neither a maximum nor a minimum value.

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

Answer: **Momentum** and **Nesterov Momentum** (also called **Nesterov** Accelerated Gradient/NAG) are slight variations of normal gradient descent that can speed up training and improve convergence significantly.

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**. ... Matrix multiplication is the essential math operation of a **neural network**.

1. What is internal covariance shift in Neural Networks?

Answer: An **internal covariate shift** occurs when there is a change in the input distribution to our network. When the input distribution changes, hidden layers try to learn to adapt to the new distribution. This slows down the training process.