**DEEP LEARNING – WORKSHEET 5**

1. Which of the following are advantages of batch normalization?

Answer: D) All of the above

2. Which of the following is not a problem with sigmoid activation function?

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

3. Which of the following is not an activation function?

Answer: D) None of the above

4. 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?’

Answer: A) True

5. In which of the weights initialisation techniques, does the variance remains same with each passing layer?

Answer: A) Bias initialisation

6. Which of the following is main weakness of AdaGrad?

Answer: A) learning rate shrinks and becomes infinitesimally small

7. In order to achieve right convergence faster, which of the following criteria is most suitable?

Answer: C) momentum and learning rate both must be low

8. When is an error landscape is said to be poor(ill) conditioned?

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

9. Which of the following Gradient Descent algorithms are adaptive?

Answer: A) ADAM C) NADAM

10. When should an optimization function (gradient descent algorithm) stop training:

Answer: D) when it reaches a local minima which is similar to global minima (i.e. which has very less error distance with global minima)

11. 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. convex optimization include the optimization of biconvex, pseudo-convex, and quasiconvex functions. Extensions of the theory of convex analysis and iterative methods for approximately solving non-convex minimization problems occur in the field of generalized convexity, also known as abstract convex analysis.

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

Answer: The optimization problem in low dimensions vs high dimensions. In low dimensions, it is true that there exists lots of local minima. However in high dimensions, local minima are not really the critical points that are the most prevalent in points of interest. When we optimize neural networks or any high dimensional function, for most of the trajectory we optimize, the critical points(the points where the derivative is zero or close to zero) are saddle points. Saddle points, unlike local minima, are easily escapable.

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

Answer: The difference between NAG and classical momentum is that NAG puts more weight on recent gradients: in fact, it gives zero weight to the first gradient descent direction after the first iteration, so the second step also be a pure gradient descent step, but with an extra big step size of (1+μ)ε. Another way of looking at it is to say that NAG forgets old gradients more quickly.

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

Answer: Weight initialization is to prevent layer activation outputs from exploding or vanishing during the course of a forward pass through a deep neural network. If either occurs, loss gradients will either be too large or too small to flow backwards beneficially, and the network will take longer to converge, if it is even able to do so at all.

15. What is internal covariance shift in Neural Networks?

Answer: Batch Normalization (BN) techniques have been proposed to reduce the so-called Internal Covariate Shift (ICS) by attempting to keep the distributions of layer outputs unchanged. Experiments have shown their effectiveness on training deep neural networks. However, since only the first two moments are controlled in these BN techniques, it seems that a weak constraint is imposed on layer distributions and furthermore whether such constraint can reduce ICS is unknown.