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?
   A) Reduces internal covariant shift.
   B) Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
   C) allows use of saturating nonlinearities and higher learning rates.
   ➡ D) All of the above
2. Which of the following is not a problem with sigmoid activation function?
   A) Sigmoids do not saturate and hence have faster convergence
   B) Sigmoids have slow convergence.
   ➡ C) Sigmoids saturate and kill gradients.
   D) Sigmoids are not zero centered; gradient updates go too far in different directions, making optimization more difficult.
3. Which of the following is not an activation function?
   A) Swish B) Maxout
   C) SoftPlus ➡ 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?
   ➡ A) True B) False
5. In which of the weights initialisation techniques, does the variance remains same with each passing layer?
   ➡ A) Bias initialisation B) Xavier Initialisation
   C) He Normal Initialisation D) None of these
6. Which of the following is main weakness of AdaGrad?
   ➡ A) learning rate shrinks and becomes infinitesimally small
   B) learning rate doesn't shrink beyond a point
   C) change in learning rate is not adaptive
   D) AdaGrad adapts updates to each individual parameter
7. In order to achieve right convergence faster, which of the following criteria is most suitable?
   ➡ A) momentum and learning rate both must be high
   B) momentum must be high and learning rate must be low
   C) momentum and learning rate both must be low
   D) momentum must be low and learning rate must be high

8. When is an error landscape is said to be poor(ill) conditioned?
   A) when it has many local minima
   B) when it has many local maxima
   ➡ C) when it has many saddle points and flat areas
   D) None of these

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

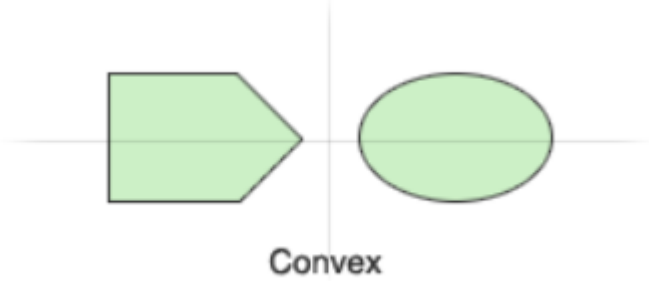
9. Which of the following Gradient Descent algorithms are adaptive?
   ➡ A) ADAM ➡ B) SGD
   ➡ C) NADAM D) RMS Prop.

10. When should an optimization function (gradient descent algorithm) stop training:
    A) when it reaches local minimum B) when it reaches saddle point
    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.

11. **What are convex, non-convex optimization?**

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. Conic optimization problems -- the natural extension of linear programming problems -- are also convex problems. In a convex optimization problem, the feasible region -- the intersection of convex constraint functions -- is a convex region, as pictured below.



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

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.

A typical problem for both local minima and saddle-points is that they are often surrounded by plateaus of small curvature in the error. While gradient descent dynamics are repelled away from a saddle point to lower error by following directions of negative curvature, this repulsion can occur slowly due to the plateau.

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

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).

14. **What is pre-initiliazation of weights?**

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. 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.

Matrix multiplication is the essential math operation of a neural network. In deep neural nets with several layers, one forward pass simply entails performing consecutive matrix multiplications at each layer, between that layer's inputs and weight matrix. The product of this multiplication at one layer becomes the inputs of the subsequent layer, and so on and so forth.

15. **What is internal covariance shift in neural networks?**

Internal Covariate Shift as the change in the distribution of network activations due to the change in network parameters during training. 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.