1. D) All of the above
2. A) Sigmoids do not saturate and hence have faster convergence
3. D) None of the above
4. A) True
5. B) Xavier Initialisation
6. A) learning rate shrinks and becomes infinitesimally small
7. B) momentum must be high and learning rate must be low
8. C) when it has many saddle points and flat areas
9. A) ADAM & D) RMS Prop.
10. A) when it reaches local minimum & 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)

11.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, With a convex objective and a convex feasible region, there can be only one optimal solution, which is globally optimal.  Several methods -- notably Interior Point methods -- will either find the globally optimal solution, or prove that there is no feasible solution to the problem.  Convex problems can be solved efficiently up to very large size.

A non-convex optimization problem is any problem where the objective or any of the constraints are non-convex

Such a problem may have multiple feasible regions and multiple locally optimal points within each region.  It can take time exponential in the number of variables and constraints to determine that a non-convex problem is infeasible, that the objective function is unbounded, or that an optimal solution is the "global optimum" across all feasible regions

12.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.The intuition with the saddle point, is that, for a minima located close to the global minima, all directions should be climbing upward; going further downward is not possible. Local minima exist, but are very close to global minima in terms of objective functions, and theoretical results suggest that some large functions have their probability concentrated between the index (the critical points) and the objective function. The index is the fraction of directions moving downward; for all values of index not 0 or 1 (local minima and maxima, respectively), then it is a saddle point.

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

14. If one neuron contains a weights vector that represents what a neuron has learnt that is multiplied with an input vector on new data;

And if the learning process is cyclical, feeding forward all data through the network.

So,it must start somewhere. And indeed, it starts at epoch 0 – or, put simply, at the start. And given the fact that during that first epoch, we’ll see a forward pass, the network cannot have empty weights whatsoever. They will have to be initialized.

In short, weight initialization comprises setting up the weights vector for all neurons for the first time, just before the neural network training process starts. As you can see, indeed, it is highly important to neural network success: without weights, the forward pass cannot happen, and so cannot the training process.

15. We define 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. When the parameters of a layer change, so does the distribution of inputs to subsequent layers.

These shifts in input distributions can be problematic for neural networks, especially deep neural networks that could have a large number of layers.