1. D
2. C
3. B,C
4. A
5. B
6. B
7. D
8. C
9. A,B
10. A,C,D
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.

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

1. 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. The thing with saddle points is that they are a type of optimum which combines a combination of minima and maxima. Because the number of dimensions is so large with deep learning, the probability that an optimum only consists of a combination of minima is very low. This means 'getting stuck' in a local minimum is rare. At the risk of oversimplifying, it's harder to 'get stuck' in a saddle point because you can 'slide down one of the dimensions
2. The difference between Momentum method and Nesterov Accelerated Gradient is in gradient computation phase. In Momentum method, the gradient was computed using current parameters θ𝑡

whereas in Nesterov Accelerated Gradient, we apply the velocity vt to the parameters θ to compute interim parameters θ̃ . We then compute the gradient using the interim parameters

1. NA
2. We define Internal Covariate Shift as the change in the distribution of network activations due to the change in network parameters during training, 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.