Q.2.

Sudha Devi
Roll No: 2020900003

2. The computational cost of stochastic gradient descent is higher than that of mini-batch gradient descent as the mini-batch gradient descent requires summing over a few examples.

Sol:- Gradient descent update for a learning rate η

Given

$$w^{k+1} \longleftarrow w^{k} - \eta \nabla J$$

Here, we want to bound ∇J

when we start at some $f(w^0)$ at each step we decrease f by at least $\frac{1}{2L}\|\nabla f(w^k)\|^2$

we cant decrease $f(w)^k$ below f^*

so, $\|\nabla f(w^k)\|^2$ must be going to 0 fast enough.

$$f(w^{k+1}) \leq f(w^k) - \frac{1}{2L}\|\nabla f(w^k)\|^2$$

Here $\frac{1}{2L}\|\nabla f(w^k)\|_2^2$ will always be positive unless $\nabla f(x) = 0$, this inequality implies that the objective function value strictly decreases with each iteration of gradient descent

until it reaches the optimal value $f(x) = f(x^*)$

This is why the gradient descent diverges when the step size is too large.

5) Effect of learning rate →

When rate is high the gradient descent changes are high and it can overshoot the convergence point

If step size/learning rate is low it will take lot of iterations to converge. So this process is slow.