Q2-b

Stochastic gradient decent is little different. gradient decent iterates over all Points and for lage data that is very Time consuming.

In SGD, we shuffle the data ; Pick Random k data Points of them and run the gradient steps. In SGD, we also have the learning rate λ. the gradient step is similar to Q2-a ($Fdx(\theta)$).

Procedure SGD (Fdx, $\theta(0)$)

$\theta \leftarrow \theta(0)$

while not converged do

for shuff data $\{1,2,\cdots N\}$ Pick K data Points

For $k \in \{1,2,\cdots K\}$ do

$$\theta_K \leftarrow \theta_K + \lambda \, Fdx(\theta_k)$$

return θ_K