**Stochastic Gradient Descent**

|  |
| --- |
| **Goal:** train massive data set |
| **What is the difference**?, theta for all training examples are not updated at once, instead, it change theta based on partial training examples, and use changed theta to predict next partial training examples. |

**Mini-batch Gradient Descent**

|  |
| --- |
| **Stochastic Gradient Descent vs Mini-batch Gradient Descent**  Stochastic gradient descent uses 1 example in each iteration  Mini-batch gradient descent use b examples in each iteration |
|  |

**Check convergence of stochastic Gradient Descent**

|  |
| --- |
| Check the cost of a single training example before updating parameters based on it.  For every 1000 iterations, average the cost for those examples before their update. If the average cost continues to decrease, the algorithm is converging.  **Note**   1. Changing the learning rate may help to converge to other local optimum 2. Large alpha may result in divergence 3. Increase the #iteration to average can smooth the learning curve 4. Slowly decrease the alpha overtime will help the algorithm to converge to global minimum |

Online Learning

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |