1 Batch Gradient Descent

1.1 Definition

Batch Gradient Descent computes the gradient of the cost function by averaging over the entire training dataset.

1.2 Advantages

- **Less noise and randomness since the entire dataset is used**:

1.3 Disadvantages

- **Computationally Intensive (Take a lot of time)**
- **High Memory Usage**

2 Stochastic Gradient Descent (SGD)

2.1 Definition

Stochastic Gradient Descent updates the parameters using the gradient computed from a single randomly selected data point at each iteration.

2.2 Advantages

- **Faster and use less memory Than Batch gradient**
- **Since it is random it may escape local minima and find the global minima**

2.3 Disadvantages

- **It may not find the global minima**

3 Mini-Batch Gradient Descent

3.1 Definition

Mini-Batch Gradient Descent updates parameters based on the gradient computed from a small batch of the data.

3.2 Advantages

- **Ability to paralleled processing**
- **more stable than SGD**

3.3 Disadvantages

- **Complex**
- **it may still need a lot of memory**

4 Momentum

4.1 Definition

Momentum gradient descent accelerates the optimization process by adding a fraction of the previous update vector to the current update.

4.2 Advantages

- **Faster convergence than standard gradient descent**
- **Reduces noise and helps escape local minima**

4.3 Disadvantages

- **May not find the global minimum if the momentum term is too large**