types of gradient descent

1. Batch gradient descent :

- the entire dataset is used to compute the loss function with respect to parameter, in each iteration, the algorithm calculates the loss function for all training examples and updates.

- pros:

*Global view: considers the entire dataset for each update, providing a global perspective for the loss landscape

*Stable Convergence**:** Since the gradient is averaged over all training examples, the updates are less noisy and more stable.

-cons:

*computationally expensive

*Memory intensive

2. Stochastic gradient descent:

- Model parameters are updated using gradient of loss function with respect to single training example at each iteration

-Pros:

*Faster convergence than that of Batch gradient descent.

*less memory intensive than Batch gradient descent.

*better for online learning : where data comes in streams allowing the data to be frequently updated.

-Cons:

*Hyperparameters sensitivity

*Potential for Overshooting**:** The frequent updates can cause the algorithm to overshoot the minimum, especially with a high learning rate.

3. Mini-batch gradient descent:

-A compromise between Batch gradient descent and stochastic gradient descent, it updates the parameters using a small random subset of training data called mini-batch

-Pros:

*Faster Convergence**:** By using mini-batches, it achieves a balance between the noisy updates of stochastic gradient descent and the stable updates of Batch Gradient Descent.

*Reduced Memory Usage**:** Requires less memory than Batch Gradient Descent as it only needs to store a mini-batch at a time.

-Cons:

*Potential for Suboptimal Mini-Batch Sizes**:** Selecting an inappropriate mini-batch size can lead to suboptimal performance and convergence issues.

*Complexity in Tuning**:** Requires careful tuning of the mini-batch size and learning rate to ensure optimal performance.

4. Momentum-Based Gradient Descent:

-an enhancement of the standard gradient descent algorithm that aims to accelerate convergence, particularly in the presence of high curvature, small but consistent gradients, or noisy gradients.

-Pros:

*Accelerated Convergence**:** Helps in faster convergence, especially in scenarios with small but consistent gradients.

*Smoother Updates: Reduces the oscillations in the gradient updates, leading to a smoother and more stable convergence path.

-Cons:

*Additional Hyperparameter**:** Introduces an additional hyperparameter (momentum term) that needs to be tuned.

*Complex Implementation**:** Slightly more complex to implement compared to standard gradient descent.

5. RMSProp (Root Mean Square Propagation):

-modifies the accumulation of squared gradients to use a moving average.This prevents the learning rate from decaying too quickly.

-Pros:

*Controlled Learning Rate Decay: Uses a moving average to prevent the learning rate from decaying too quickly.

*Adaptable to Non-Stationary Objectives**:** More suitable for non-stationary objectives due to the moving average mechanism.

-Cons:

*Potential for Oscillation**:** If the decay rate is not properly tuned, it can lead to oscillations in the learning process.

*Hyperparameter Sensitivity