**Stochastic Gradient Descent (SGD)**

* Gradient Descent is an iterative process to find the optimum value (minimum or maximum) of an objective function.
* It is one of the most used methods for changing a model’s parameters in order to reduce a cost function in machine learning models.
* The primary goal of gradient descent is to identify the model parameters that provide the maximum accuracy on both training and test datasets.
* The algorithm might gradually drop towards lower values of the function by moving in the opposite direction of the gradient, until reaching the minimum of the function.

**Types of Gradient Descent :**

Typically, there are three types of Gradient Descent:

* [Batch Gradient Descent](https://www.geeksforgeeks.org/difference-between-batch-gradient-descent-and-stochastic-gradient-descent/)
* Stochastic Gradient Descent
* [Mini-batch Gradient Descent](https://www.geeksforgeeks.org/ml-mini-batch-gradient-descent-with-python/)

## Stochastic Gradient Descent (SGD):

* Stochastic Gradient Descent (SGD) is a variant of the [Gradient Descent](https://www.geeksforgeeks.org/gradient-descent-algorithm-and-its-variants/)algorithm that is used for optimizing [machine learning](https://www.geeksforgeeks.org/machine-learning-algorithms/) models.
* It overcomes the computational inefficiency of traditional Gradient Descent methods, making it suitable for large datasets in machine learning.
* In SGD, instead of using the entire dataset for each iteration, only a single random training example (or a small batch) is selected to calculate the gradient and update the model parameters.
* This random selection introduces randomness into the optimization process, hence the term “stochastic” in stochastic Gradient Descent

**Stochastic Gradient Descent Algorithm**

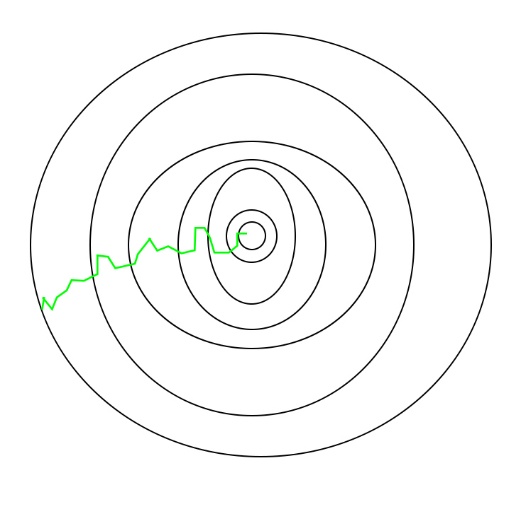
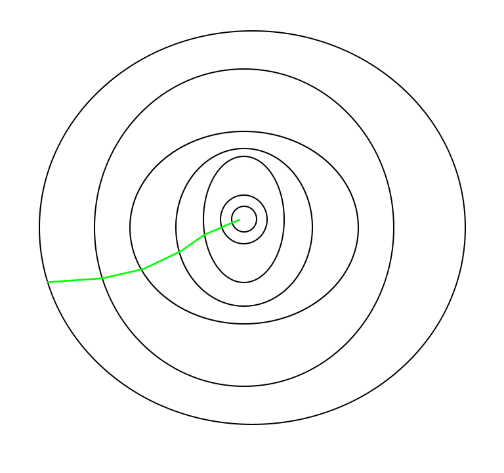
1. Initialization: Randomly initialize the parameters of the model.
2. Set Parameters: Determine the number of iterations and the learning rate (alpha) for updating the parameters.
3. Stochastic Gradient Descent Loop: Repeat the following steps until the model converges or reaches the maximum number of iterations:

* Shuffle the training dataset to introduce randomness.
* Iterate over each training example (or a small batch) in the shuffled order.
* Compute the gradient of the cost function for the current example/batch.
* Update model parameters by stepping in the negative gradient direction, scaled by the learning rate.
* Check convergence criteria (e.g., cost function difference).

1. **Return Optimized Parameters:** Output the optimized parameters after meeting criteria or completing iterations.

SGD uses a single data point (or small batch) per iteration, resulting in a noisier but faster path to the minimum compared to traditional Gradient Descent.

**The path taken by Batch Gradient Descent | Stochastic Gradient Descent looks as follows**



SGD, being noisier than traditional Gradient Descent, typically requires more iterations to reach the minima due to its randomness.

However, it remains computationally less expensive, making it the preferred choice over Batch Gradient Descent in most scenarios for optimizing learning algorithms.

**Advantages of Stochastic Gradient Descent (SGD):**

1. **Speed:** Faster than Batch and Mini-Batch Gradient Descent, as it updates parameters using a single example.
2. **Memory Efficiency:** Handles large datasets efficiently since it processes one example at a time.
3. **Avoidance of Local Minima:** Noisy updates enable SGD to escape local minima and move toward the global minimum.

**Disadvantages of Stochastic Gradient Descent (SGD):**

1. **Noisy Updates:** High variance in updates can cause instability and oscillations around the minimum.
2. **Slow Convergence:** Requires more iterations since updates are based on individual examples.
3. **Sensitivity to Learning Rate:** A poorly chosen learning rate can cause overshooting or slow convergence.
4. **Less Accurate:** May not reach the exact global minimum due to noise, though this can be mitigated with techniques like learning rate scheduling and momentum.

**Comparison: Stochastic Gradient Descent (SGD) vs. Batch Gradient Descent**

| **Aspect** | **Stochastic Gradient Descent (SGD)** | **Batch Gradient Descent** |
| --- | --- | --- |
| **Dataset Usage** | Processes a single random sample or small batch per iteration. | Processes the entire dataset per iteration. |
| **Computational Efficiency** | Computationally less expensive per iteration. | More expensive per iteration. |
| **Convergence** | Faster convergence due to frequent updates. | Slower convergence due to infrequent updates. |
| **Noise in Updates** | High noise due to small sample updates. | Low noise with updates based on all data. |
| **Stability** | Less stable; may oscillate around the optimum. | More stable; converges smoothly. |
| **Memory Requirement** | Requires less memory. | Requires memory to store the entire dataset. |
| **Update Frequency** | Frequent updates; ideal for online learning and large datasets. | Infrequent updates; better for smaller datasets. |
| **Initialization Sensitivity** | Less sensitive to initial parameters. | More sensitive to initial parameters. |