**Vietnam General Confederation of Labor**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**FINAL EXAM REPORT**

**MACHINE LEARNING**

*Instructor*: **MR. LE ANH CUONG**

*Student*: **Nguyen Pham Thanh Uyen – 521H0329**

*Class* : **21H50302**

*Year* ***:* 2023-2024**

**HO CHI MINH CITY, 2023**

Vietnam General Confederation of Labor

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**FINAL EXAM REPORT**

**MACHINE LEARNING**

*Instructor*: **MR. LE ANH CUONG**

*Student*: **Nguyen Pham Thanh Uyen – 521H0329**

*Class* **: 21H50302**

*Year* **: 2023-2024**

**HO CHI MINH CITY, 2023**

ACKNOWLEDGEMENT

First, I would like to send my most sincere thanks to Mr. Le Anh Cuong, who enthusiastically taught me to help me gain full knowledge of the subject and complete this report.

Second, I would like to send my thanks to the teachers of the Department of Information Technology at Ton Duc Thang University for giving me the opportunity to write this report.

I look forward to receiving feedback from teachers so I can improve my report.

Finally, I would like to wish the teachers good health and success in their careers.

*Ho Chi Minh city, 21th December 2023*

*Author*

*(Sign and write full name)*

*Nguyen Pham Thanh Uyen*

**THIS PROJECT WAS COMPLETED AT**

**TON DUC THANG UNIVERSIY**

I fully declare that this is my own project and is guided by Mr. Le Anh Cuong; The research contents and results in this topic are honest and have not been published in any form before. The data in the tables for analysis, comments and evaluation are collected by the author himself from different sources, clearly stated in the reference section.

Besides that, the project also uses several comments, assessments as well as data from other authors, other agencies and organizations, with citations and source annotations.

**Should any frauds be found, I will take full responsibility for the content of my report.** Ton Duc Thang University is not related to copyright and copyright violations caused by me during the implementation process (if any).

*Ho Chi Minh city, 21th December, 2023*

*Author*

*(Sign and write full name)*

*Nguyen Pham Thanh Uyen*

CONFIRMATION AND ASSESSMENT SECTION

**Instructor confirmation section**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh December 2023*

*(Sign and write full name)*

**Evaluation section for grading instructor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh December 2023*

*(Sign and write full name)*

SUMMARY

This is the final report on Machine Learning of the Faculty of Information Technology of Ton Duc Thang University.

My writing will sometimes have many errors. I am very open to receiving constructive contributions from teachers and will take them as lessons to improve.

INDEX

[ACKNOWLEDGEMENT i](#_Toc154261618)

[CONFIRMATION AND ASSESSMENT SECTION iii](#_Toc154261619)

[SUMMARY iv](#_Toc154261620)

[INDEX 1](#_Toc154261621)

[CHAPTER 1 – LEARN AND COMPARE OPTIMIZER METHODS IN TRAINING MACHINE LEARNING MODELS 2](#_Toc154261622)

[1. Learn Optimizer methods in training machine learning models. 2](#_Toc154261623)

[1.1 Gradient Descent (GD) 2](#_Toc154261624)

[1.2 Stochastic Gradient Descent (SGD) 6](#_Toc154261625)

[1.3 Momentum 8](#_Toc154261626)

[1.4 AdaGrad 9](#_Toc154261627)

[2. Compare Optimizer methods in training machine learning models. 11](#_Toc154261628)

[CHAPTER 2 – LEARN ABOUT CONTINUOUS LEARNING AND TEST PRODUCTION WHEN BUILDING A MACHINE LEARNING SOLUTION TO SOLVE A CERTAIN PROBLEM 14](#_Toc154261629)

[CHAPTER 3 – REFERENCES 16](#_Toc154261630)

CHAPTER 1 – LEARN AND COMPARE OPTIMIZER METHODS IN TRAINING MACHINE LEARNING MODELS

1. Learn Optimizer methods in training machine learning models.

Basically, the optimization algorithm is the basis for building a neural network model with the purpose of ‘learning’ the features (or pattern) of input data, from which a suitable pair of weights and bias can be found to optimize the model.

The reason optimization algorithms exist is to improve the weights and bias step by step.

* 1. **Gradient Descent (GD)**

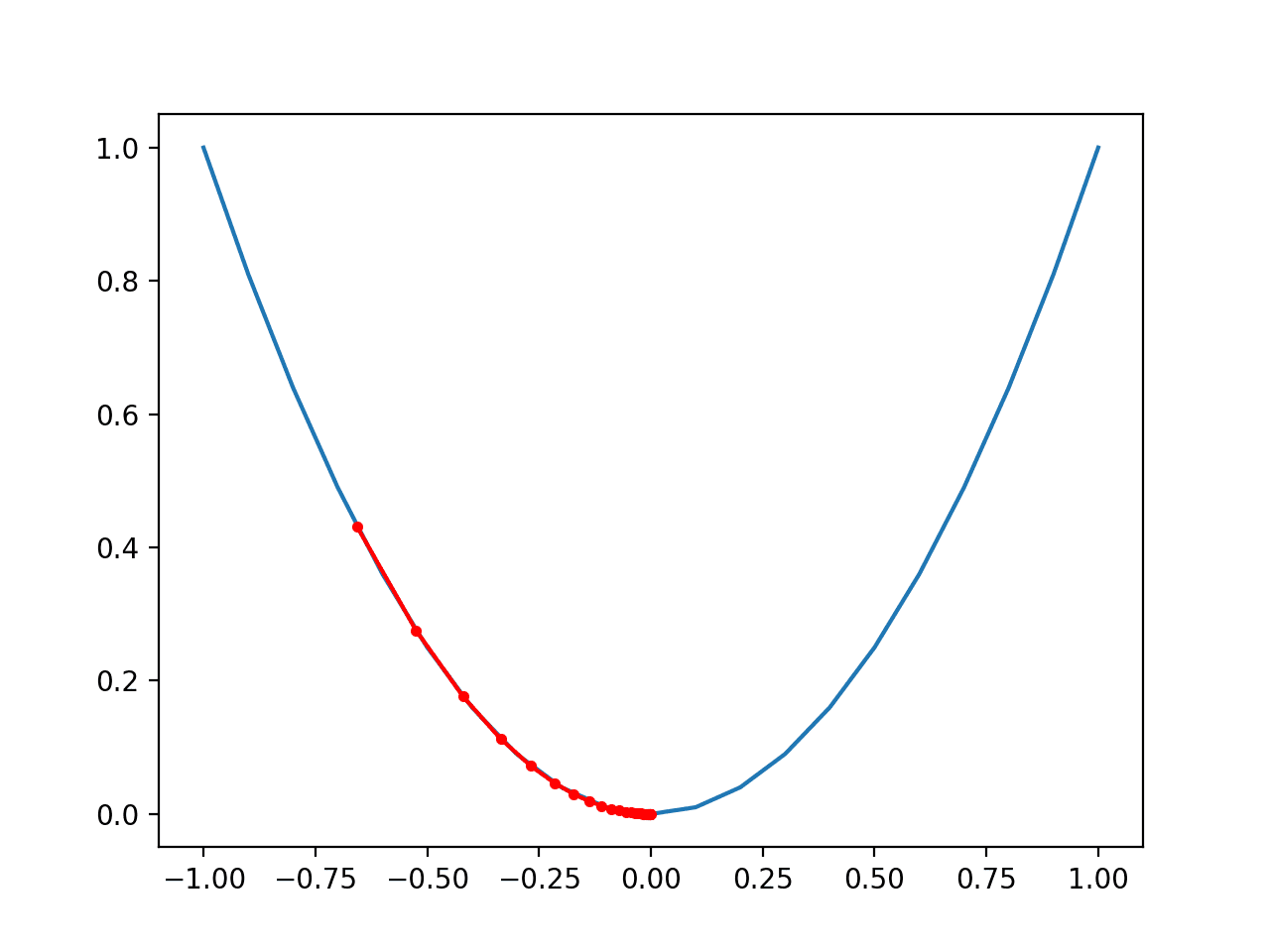
Gradient descent is an optimization algorithm used to minimize the cost function of a model. The cost function measures how well the model fits the training data and is defined based on the difference between the predicted and actual values.

Gradient: Rate of inclination or declination of a slope.

Descent: Descent: Is the abbreviation for descending, meaning gradually decreasing.

* + 1. **Gradient Descent for a function of one variable**

Consider the function of one variable . The Gradient Descent algorithm finds the minimum of a function by initializing the value at a random position, then moving in the opposite direction to the derivative of . This operation will be repeated until a certain threshold is reached.



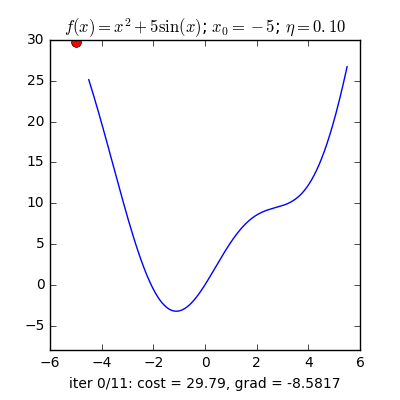
Update formula :

In there:

: value at step .

: Learning rate.

: Derivative of function at .



Ảnh có chứa văn bản, Sơ đồ, biểu đồ, hàng

Mô tả được tạo tự động

*Illustration of the Gradient Descent algorithm for a function of one variable*

* + 1. **Gradient Descent for multi-variable functions**

For a multivariable function , this algorithm will calculate the Gradient (derivative vector) of the function at a random point , then move in the opposite direction to this Gradient.

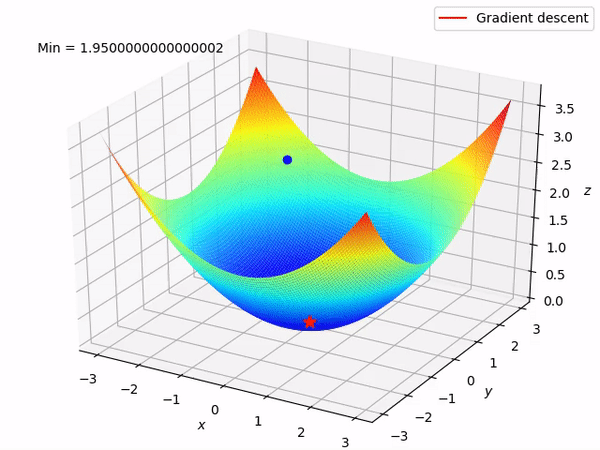
Recipe:

In there:

: value at step .

: Learning rate.

: Gradient of function at .



*Illustration of Gradient Descent algorithm for multi-variable function*

* + 1. **Advantages and disadvantages of Gradient Descent**

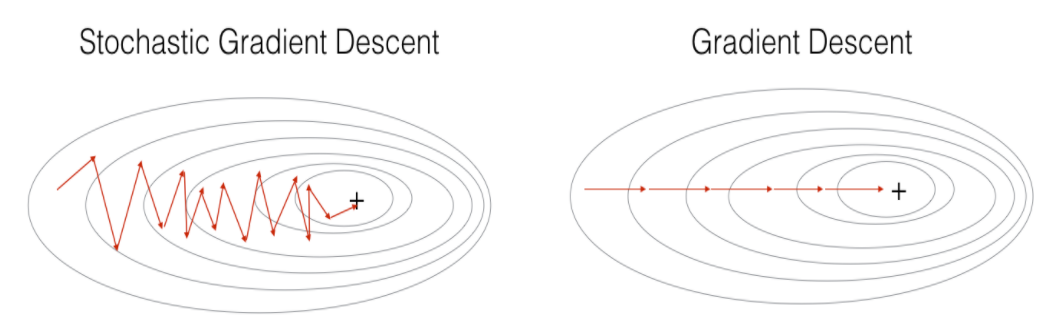
***Advantages:***

* Gradient Descent is one of the most popular and widely used optimization methods in machine learning.
* Basic Gradient Descent algorithm, easy to understand.
* Solve neural network model optimization problems by updating weights after each loop.

***Disadvantages:***

* Because of its simplicity, the Gradient Descent algorithm has many limitations such as depending on the initial solution and learning rate.
* A learning rate that is too large may cause the algorithm to fail to converge, while a learning rate that is too small may slow down the learning process.
* If the data has large noise, Gradient Descent may converge unevenly or be affected by noisy data.
  1. **Stochastic Gradient Descent (SGD)**

Stochastic is a variation of Gradient Descent. Instead of updating the weight once after each epoch, in each epoch with N data points we will update the weight N times. On the one hand, SGD will reduce the speed of one epoch. However, looking in another direction, SGD will converge very quickly after only a few epochs. The SGD formula is similar to GD but is performed on each data point.



Looking at the two pictures above, we see that SGD has a rather zigzag path, not as smooth as GD. That's because one data point cannot represent the entire data. Here, GD has limitations for large databases (several million data), so calculating the derivative on the entire data through each loop becomes cumbersome. Besides, education is not suitable for online learning because the calculation time is long, and the algorithm is no longer online. Therefore, SGD was born to solve that problem, because each time new data is added, only one data point needs to be updated, suitable for online learning.

* + 1. **Advantages and disadvantages of Stochastic Gradient Descent**

***Advantages:***

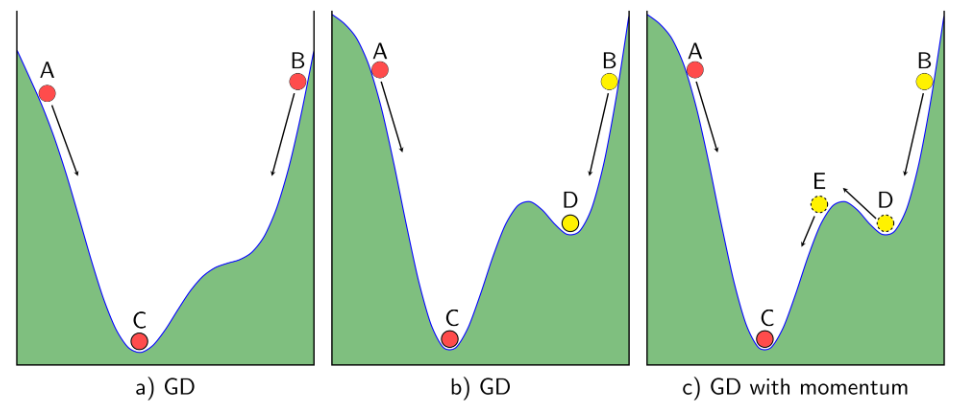
* SGD is suitable for large data processing as it only requires one data sample to update parameters, significantly reducing memory usage.
* SGD is flexible and it can be easily integrated into complex machine learning models and other optimization algorithms.
* SGD can minimize the influence of noise in the data due to randomness during the update process.

***Disadvantages:***

* The algorithm has not yet solved the two major disadvantages of GD: learning rate and initial data points.
* Due to randomness, SGD may get stuck at the local minimum and fail to reach the global minimum.
  1. **Momentum**

To overcome the above limitations of the Gradient Descent algorithm, people use gradient descent with momentum.

Momentum is a technique used in the optimization process of optimization algorithms such as Gradient Descent to help improve the convergence performance and stability of the optimization process.



Looking from a mathematical perspective, we have the Momentum vector formula:

In there:

: Momentum vector at time step .

: Momentum coefficient, usually set between 0 and 1.

Then we use the momentum vector to update the parameter.

In there:

: Parameter at time step +1.

: Learning rate, determines the size of the update step.

* + 1. **Advantages and disadvantages of Momentum**

***Advantages:***

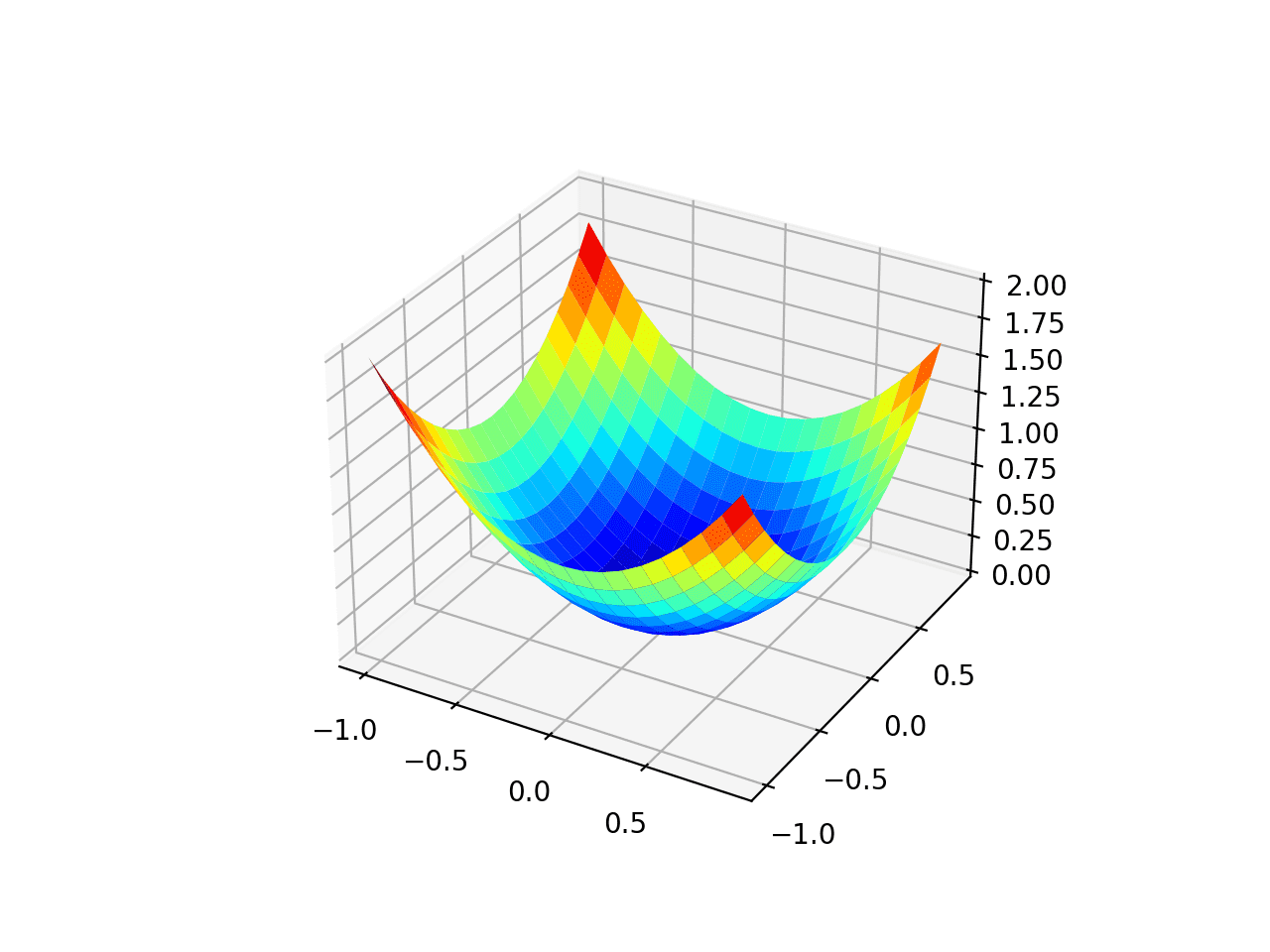
* The optimal algorithm solves the problem: Gradient Descent does not reach the global minimum point but only stops at the local minimum.

***Disadvantages:***

* Momentum can also increase the likelihood of "exceeding" the local optimum and may need to be combined with other strategies such as learning rate schedules to achieve best results.
  1. **AdaGrad**

The Adaptive Gradient Algorithm, or AdaGrad for short, is an extension of the gradient descent optimization algorithm.

It is designed to speed up the optimization process, such as reducing the number of function evaluations needed to reach the optimum, or to improve the ability of the optimization algorithm, e.g. to yield the result same better.



In there:

: Constant.

: Gradient at time .

: Error avoidance factor (divided by sample equal to 0).

: Diagonal matrix where each element on the diagonal (i,i) is the square of the parameter vector derivative at time .

* + 1. **Advantages and disadvantages of AdaGrad**

***Advantages:***

* AdaGrad has the advantage of avoiding adjusting the learning rate manually. Just set the default learning rate to 0.01 and the algorithm will automatically adjust.
* Efficient in processing multidimensional data.

***Disadvantages:***

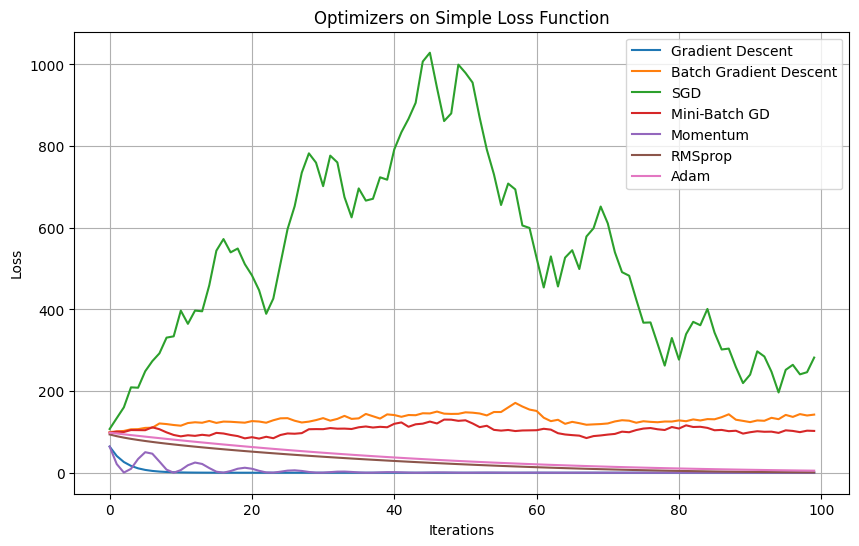
* Because the square of the gradient accumulates, the learning rate decreases over time, which can slow down the convergence process.
* Parameters with large gradients are often updated quickly, which can lead to too large a learning rate.

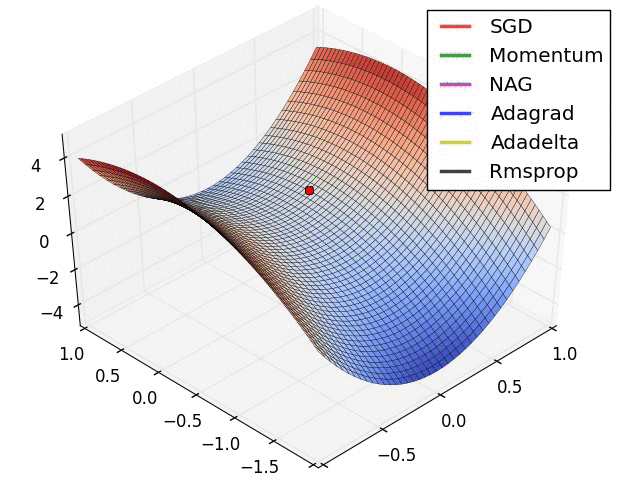
**2. Compare Optimizer methods in training machine learning models.**

**2.1 Compare using table:**

|  |  |  |
| --- | --- | --- |
| **METHOD** | **EFFICIENCY** | **CONVERGENCE SPEED** |
| Gradient Descent (GD) | Solve small and simple problems well. | Can get stuck at local minima without corrective techniques. |
| Stochastic Gradient Descent (SGD) | Effective with large and heterogeneous data sets. | Highly volatile and can be unstable due to randomness. |
| Momentum | Reduces oscillations and provides faster convergence on highly oscillatory loss function surfaces.  Effective in passing through the minimum point. | Parameters need to be adjusted to achieve best performance. |
| AdaGrad | Efficient in handling high-dimensional data. | Excessive updates for frequently updated parameters. |

**2.2 Compare graphically on simple and complex problem types:**





CHAPTER 2 – LEARN ABOUT CONTINUOUS LEARNING AND TEST PRODUCTION WHEN BUILDING A MACHINE LEARNING SOLUTION TO SOLVE A CERTAIN PROBLEM

**2.1 Continuous learning**

**2.1.1** **The concept of continuous learning**

Continuous learning is an area of machine learning in which the learning model must continuously learn and update its knowledge over time, regularly and automatically based on new data. This is necessary to ensure that the model always reflects changes in the data and environment.

Continuous learning faces two main challenges: forgetting old knowledge and interfering with already learned knowledge (the phenomenon of "catastrophic forgetting"). To solve this problem, the Continuous Learning method focuses on developing mechanisms to retain old knowledge and control the learning of new knowledge.

2.1.2 Mechanism of Continuous Learning

* Replay-based methods

In this method, historical data is "repeated" to train the model on old and new data. Old data is stored and reused as an expanded training set.

*Advantages:*

- Helps the model retain important information from old data.

- Make sure that the model does not completely forget what it has learned.

*Disadvantages:*

- Requires large storage to retain old data.

- May not be effective if the old data set is too large.

* Regularization-based Methods:

Regularization is the addition of regularization terms to the loss function to stabilize and control the learning process. In the context of continuous learning, regularization can be designed to ensure the model does not forget too much information from old data.

*Advantages:*

- Increase the generality of the model.

* Dynamic Architecture Methods:

This method focuses on changing the model's architecture over time to adapt to changes in data or tasks. The architecture can change automatically or be controlled by a management structure.

*Advantages:*

- Allows the model to flexibly adapt to changes in data and tasks.

- Helps reduce model complexity.

* Application of Continuous Learning:

Continuous Learning can be applied in many situations in building machine learning solutions.

CHAPTER 3 – REFERENCES