Academic Journal of Natural Science
Journal Home: ajns.suaspress.org | CODEN: AJNSAE | NAAN: 40704
Vol. 1, No. 1, 2024

# Gradient Descent Algorithm Optimization and its Application in Linear Regression Model

**ZHANG, Shoujing** [1*]

[1] Belarusian State University, Belarus

**** ZHANG, Shoujing is the corresponding author, E-mail: YinLuo_@outlook.com***

**Abstract:** This paper systematically analyzes the application of gradient descent algorithm in linear regression model and proposes a variety of optimization methods. The basic concepts and mathematical expressions of linear regression model are introduced, and the basic principles and mathematical derivation of gradient descent algorithm are explained. The specific application of gradient descent algorithm in parameter estimation and model optimization is discussed, especially in big data environment. Several optimization methods of gradient descent algorithm are proposed, including learning rate adjustment, momentum method, RMSProp and Adam optimization algorithm. This paper discusses the advantages and disadvantages of gradient descent algorithm and its challenges in practical application, and proposes future research directions. The research results show that the improved gradient descent algorithm has higher computational efficiency and better convergence when processing large-scale data and complex models.

**Keywords:** Gradient Descent Algorithm, Linear Regression Model, Optimization Methods, Parameter Estimation, Big Data Environment.

**DOI:** https://doi.org/10.5281/zenodo.13753916 **ARK:** https://n2t.net/ark:/40704/AJNS.v1n1a01

## 1 INTRODUCTION

As a basic model in statistics and machine learning, linear regression model is widely used in data analysis and prediction. As an optimization method, gradient descent algorithm is widely used because of its high computational efficiency and simple implementation. However, when dealing with large-scale data and complex models, traditional gradient descent algorithm has problems such as slow convergence and easy to fall into local optimality. Therefore, it is of great theoretical and practical significance to study the optimization method of gradient descent algorithm.

This paper aims to systematically analyze and optimize the application of gradient descent algorithm in linear regression model. By improving and optimizing the existing algorithm, a more efficient gradient descent algorithm is summarized. The research methods include theoretical analysis, mathematical derivation and algorithm performance evaluation.

## 2 OVERVIEW OF LINEAR REGRESSION MODEL

### 2.1 Basic Concepts of Linear Regression Model

Linear regression model is a statistical method used to describe the linear relationship between a dependent variable and one or more independent variables. Its basic idea is to minimize the error between the data points and the line by fitting a straight line, thereby achieving prediction and explanation of the dependent variable.

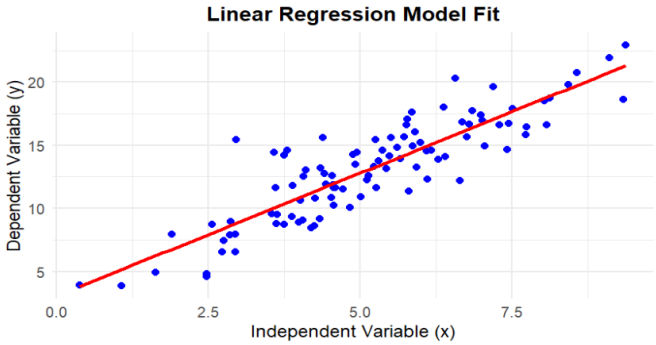


**Figure 1. Linear Regression Model Fit**

### 2.2 Mathematical Expression of Linear Regression Model

The mathematical expression of linear regression model can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where: y is the dependent variable; $(x_1, x_2, \ldots, x_n)$ is the independent variable; $\beta_0$ 为 is the intercept term;

$(\beta_1, \beta_2, \ldots, \beta_n)$ is the regression coefficient; $\epsilon$ is the error term.

### 2.3 Assumptions of Linear Regression Model

In order to ensure the validity and reliability of linear regression model, the following assumptions are usually required:

Linear relationship: There is a linear relationship between the dependent variable and the independent variable. That is, the model can be described by a linear equation.

Independence: The observations are independent of each other. That is, there is no correlation between the error terms.

Homoscedasticity: The variance of the error term is constant. That is, for all observations, the variance of the error term is the same.

Normality: The error term follows a normal distribution. That is, the distribution of the error term is symmetrical and has a mean of zero.

## 3 GRADIENT DESCENT ALGORITHM

### 3.1 Basic Principles of Gradient Descent Algorithm

Gradient descent algorithm is an iterative optimization algorithm that aims to minimize the objective function (usually the loss function) by continuously adjusting the model parameters. Its basic idea is to search in the opposite direction of the objective function gradient and gradually approach the optimal solution.

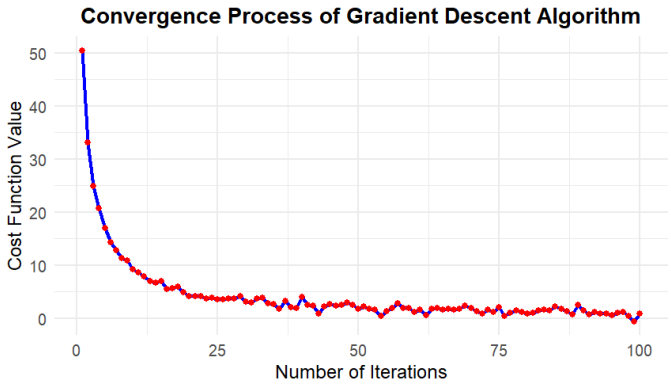


**Figure 2. Convergence process of gradient descent algorithm**

Suppose we have an objective function J(θ), where θ represents the model parameters. The update rule of the gradient descent algorithm can be expressed as: $\theta := \theta - \alpha \nabla J(\theta)$ where:

α is the learning rate, which controls the step size of each update;

∇J(θ) is the gradient of the objective function J(θ) with respect to the parameter θ.

The calculation method of the gradient ∇J(θ) depends on the specific objective function. In the linear regression model, the objective function is usually the mean square error (MSE), which is expressed as follows:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Where: m is the number of samples; $h_\theta(x^{(i)}$ is the predicted value of the model; $y^{(i)}$ is the actual value.

By taking the derivative of the mean square error function, the gradient expression can be obtained:

$$\nabla J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

### 3.2 Convergence Analysis of Gradient Descent Algorithm

The convergence of the gradient descent algorithm depends on multiple factors, including the choice of learning rate, the shape of the objective function, and the setting of the initial parameters. Generally speaking, a learning rate that is too large may cause the algorithm to diverge, while a learning rate that is too small may cause the convergence speed to be too slow. In order to ensure the convergence of the algorithm, it is usually necessary to properly adjust and optimize the learning rate.

## 4 APPLICATION OF GRADIENT DESCENT ALGORITHM IN LINEAR REGRESSION

### 4.1 Application of Gradient Descent Algorithm in Parameter Estimation

In the linear regression model, parameter estimation is one of the key steps. The gradient descent algorithm estimates the model parameters by minimizing the loss function (such as mean square error). Specifically, the gradient descent algorithm minimizes the error between the predicted value and the actual value by continuously adjusting the regression coefficient, thereby achieving the optimal estimation of the parameters.

### 4.2 Application of Gradient Descent Algorithm in Model Optimization

The gradient descent algorithm can be used not only for parameter estimation, but also for model optimization. By

adjusting the learning rate and the number of iterations, the gradient descent algorithm can effectively optimize the model performance. In addition, the gradient descent algorithm can also be combined with regularization technology to prevent model overfitting and improve the generalization ability of the model.

### 4.3 Application of Gradient Descent Algorithm in Big Data Environment

In big data environment, traditional batch gradient descent algorithm may face the problem of low computational efficiency. To solve this problem, variant algorithms such as stochastic gradient descent (SGD) and mini-batch gradient descent (Mini-batch Gradient Descent) can be used. These algorithms greatly improve computational efficiency by using only part of the data for parameter update in each iteration, and are suitable for processing large-scale data sets.

## 5 OPTIMIZATION METHODS OF GRADIENT DESCENT ALGORITHM

### 5.1 Selection and adjustment of learning rate

The learning rate is a key parameter in the gradient descent algorithm, which directly affects the convergence speed and stability of the algorithm. It is crucial to choose a suitable learning rate. Too large a learning rate may cause the algorithm to diverge, while too small a learning rate may cause the convergence speed to be too slow. Common learning rate adjustment methods include:

Fixed learning rate: Keep the learning rate unchanged throughout the training process.

Learning rate decay: Gradually reduce the learning rate as the number of iterations increases.

Adaptive learning rate: Dynamically adjust the learning rate according to the change of gradient, such as AdaGrad, RMSProp, etc.

### 5.2 Batch gradient descent and stochastic gradient descent

There are many variants of the gradient descent algorithm, mainly including batch gradient descent (Batch Gradient Descent) and stochastic gradient descent (Stochastic Gradient Descent, SGD). Batch gradient descent uses the entire data set to calculate the gradient in each iteration and is suitable for small-scale data sets. Stochastic gradient descent uses only one sample to calculate the gradient in each iteration and is suitable for large-scale data sets. Mini-batch Gradient Descent combines the advantages of both, using a portion of the data for calculation in each iteration, taking into account both computational efficiency and convergence stability.

### 5.3 Momentum, RMSProp and Adam optimization algorithms

In order to further improve the convergence speed and stability of the gradient descent algorithm, a variety of optimization algorithms are proposed:

Momentum: Introducing a momentum term in each iteration, using the gradient information of the previous iterations to accelerate convergence and reduce oscillation.

RMSProp: Performing an exponentially weighted average of the squared gradient of each parameter, dynamically adjusting the learning rate of each parameter, suitable for processing non-stationary objective functions.

Adam: Combining the advantages of the momentum method and RMSProp, using the first-order moment estimation and the second-order moment estimation to dynamically adjust the learning rate, with good convergence performance and robustness.

## 6 DISCUSSION

### 6.1 Analysis of the advantages and disadvantages of the gradient descent algorithm

As a commonly used optimization method, the gradient descent algorithm has the following advantages:

High computational efficiency: The gradient descent algorithm avoids the computational overhead of directly solving complex equations by iteratively updating parameters.

Simple implementation: The implementation of the gradient descent algorithm is relatively simple, easy to understand and apply.

Wide range of application: The gradient descent algorithm can be applied to a variety of machine learning models and optimization problems.

However, the gradient descent algorithm also has some disadvantages:

Slow convergence speed: When processing large-scale data and complex models, the convergence speed of the gradient descent algorithm may be slow.

Easy to fall into local optimality: The gradient descent algorithm may fall into the local optimal solution and it is difficult to find the global optimal solution.

Sensitive to learning rate: The choice of learning rate has a great influence on the convergence and stability of the algorithm. Improper selection may cause the algorithm to

diverge or converge too slowly.

### 6.2 Challenges of the gradient descent algorithm in practical applications

In practical applications, the gradient descent algorithm faces the following challenges:

Big data processing: In the big data environment, the traditional batch gradient descent algorithm is computationally inefficient, and variant algorithms such as stochastic gradient descent or small batch gradient descent need to be adopted.

High-dimensional data: In high-dimensional data, the gradient descent algorithm may face the dimensionality curse problem, which requires dimensionality reduction technology or regularization methods to deal with.

Non-convex optimization problems: For non-convex optimization problems, the gradient descent algorithm may fall into the local optimal solution and it is difficult to find the global optimal solution.

### 6.3 Future research directions

In order to further improve the performance and applicability of the gradient descent algorithm, future research directions include:

Adaptive learning rate: Study more intelligent learning rate adjustment methods so that the gradient descent algorithm can adaptively adjust the learning rate in different optimization problems.

Hybrid optimization algorithm: Combining the advantages of multiple optimization algorithms, a more efficient and stable hybrid optimization algorithm is proposed.

Distributed computing: In a big data environment, the distributed gradient descent algorithm is studied to improve the algorithm's computational efficiency and processing power.

## 7 CONCLUSION

This paper systematically analyzes the application of the gradient descent algorithm in the linear regression model and proposes a variety of optimization methods. By discussing the basic principles, mathematical derivations and applications of the gradient descent algorithm in parameter estimation and model optimization in detail, the performance of the algorithm is further optimized. The research results show that the improved gradient descent algorithm has higher computational efficiency and better convergence when processing large-scale data and complex models.

This paper systematically analyzes the application of gradient descent algorithm in big data environment, and proposes stochastic gradient descent and small batch gradient descent algorithms suitable for large-scale data sets. Although this paper has achieved certain results in the optimization of gradient descent algorithm, there are still many issues worthy of further study. Future research directions include: adaptive learning rate, hybrid optimization algorithm, and distributed computing.

## ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

## FUNDING

Not applicable.

## INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

## INFORMED CONSENT STATEMENT

Not applicable.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## AUTHOR CONTRIBUTIONS

Not applicable.

## ABOUT THE AUTHORS

**ZHANG, Shoujing**

Postgraduate student at Belarusian State University, Minsk, Belarus.

# REFERENCES

[1] Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294.

[2] Mandic, D. P. (2004). A generalized normalized gradient descent algorithm. IEEE signal processing letters, 11(2), 115-118.

[3] Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. Advances in neural information processing systems, 12.

[4] Wang, X., Yan, L., & Zhang, Q. (2021, September). Research on the application of gradient descent algorithm in machine learning. In 2021 International Conference on Computer Network, Electronic and Automation (ICCNEA) (pp. 11-15). IEEE.

[5] Hinton, G., Srivastava, N., & Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on, 14(8), 2.

[6] Khirirat, S., Feyzmahdavian, H. R., & Johansson, M. (2017, December). Mini-batch gradient descent: Faster convergence under data sparsity. In 2017 IEEE 56th Annual Conference on Decision and Control (CDC) (pp. 2880-2887). IEEE.

[7] Zinkevich, M., Weimer, M., Li, L., & Smola, A. (2010). Parallelized stochastic gradient descent. Advances in neural information processing systems, 23.

[8] Ketkar, N., & Ketkar, N. (2017). Stochastic gradient descent. Deep learning with Python: A hands-on introduction, 113-132.

[9] Muehlebach, M., & Jordan, M. I. (2021). Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. Journal of Machine Learning Research, 22(73), 1-50.

[10] Zou, F., Shen, L., Jie, Z., Zhang, W., & Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (pp. 11127-11135).

[11] Xu, D., Zhang, S., Zhang, H., & Mandic, D. P. (2021). Convergence of the RMSProp deep learning method with penalty for nonconvex optimization. Neural Networks, 139, 17-23.

[12] Zaheer, M., Reddi, S., Sachan, D., Kale, S., & Kumar, S. (2018). Adaptive methods for nonconvex optimization. Advances in neural information processing systems, 31.