Recent Improvements of Gradient Descent Method for Optimization

Shweta Agrawal
Professor I.A.C. S.A.G.E.
University Indore
M.P. India

Ravishek Kumar Singh
M Tech 4th Semester
Department of C.S.E. S.I.R.T
M.P. India

ABSTRACT
Gradient descent is best and common method used for optimization. Gradient descent is one of the optimization techniques apply when machine learning based model or algorithm or trained. It has a function of convex, the technique is based on this function and the parameters of this function iteratively apply to reduce cost function to find local minima. Gradient used a function which takes more than one input variable. Gradient descent technique measures the variations in the weights with respect to the change in error. The purpose of Gradient descent technique is to make changes in set of parameters for reaching optimal parameters. The purpose of Gradient descent technique is to found set that leads to the minimum loss function value possible. In this paper we introduce common optimization technique and their challenges and how this leads to the derivation by using their update rules. In this paper we give also provides advantage and disadvantage of different variants of gradient descent techniques.

Keywords
Gradient Descent, Machine learning, Optimization, cost function, Iterative

1. INTRODUCTION
Gradient descent is best and common method used for optimization. Gradient descent is one of the optimization techniques apply when machine learning based model or algorithm or trained. It has a function of convex, the technique is based on this function and the parameters of this function iteratively apply to reduce cost function to find local minima. Gradient used a function which takes more than one input variable. Gradient descent technique measures the variations in the weights with respect to the change in error. The purpose of Gradient descent technique is to make changes in set of parameters for reaching optimal parameters. The purpose of Gradient descent technique is to found set that leads to the minimum loss function value possible [1, 2].
To reach local minima gradient descent technique we need to set the rate of learning. We used an appropriate value for learning rate, which is neither very little nor very big. Selection of value for learning rate is important because if select value very big, it may be possible that we don't reach at local minima as of bounces back and sometimes the gradient descent has local minima. If select the value for rate of learning is very less, then gradient descent will ultimately reach the local minima but may be possible that may take a while[3,4].

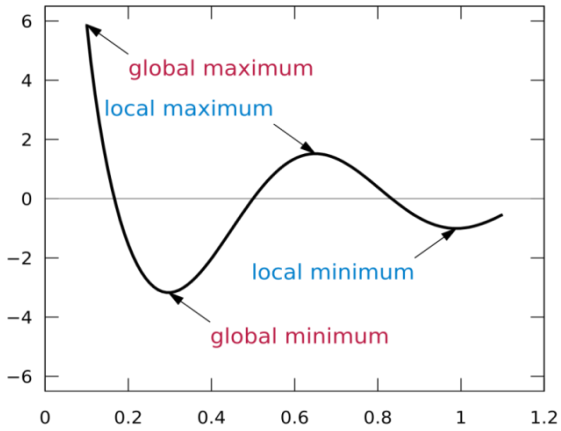

Fig 1: Maximum for local and global

2. LITERATURE SURVEY
In 2017 Shuang Song et al proposed "Stochastic gradient descent with differentially private updates". They provide a new and derive different private types of stochastic gradient, and they also tested. With the help of results they showed that standard experience has high variability because of differential privacy. They showed that reasonable growth in the batch size will increase performance significantly with a level. They showed in many cases the performance of proposed approach differentially was near to that of not private approach, specifically when we have larger sizes in batch [6].

In 2018 Loucas Pillaud et al proposed "Statistical Optimality of Stochastic Gradient Descenton Hard Learning Problems through Multiple Passes". They considered stochastic gradient technique and used it for least squares regression. They used it with possibly number of passes above the data. They showed that when there are several passes used we have seen the performance practically better. The performance can be measure in terms of predictive analysis on unseen data. They also illustrated and used synthetic data with for results and experiments. They applied kernel methods which are based on non-linear and classical standard for linear model [7].

In 2018 Prateek Jain, al proposed "Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification". They describe several benefits of averaging techniques. They showed that this technique commonly applied in conjunction with stochastic gradient techniques. The work that they presented is a sharp analysis of Mini batching approach. The method which is used for averaging provides many samples of a stochastic gradient technique. They used it for both reduce the variance and estimate and for parallelizing of a stochastic

gradient technique [9].

In 2019 Jonathan et al proposed "A Recent advances and applications of machine learning in solid state materials science". They describe a comprehensive and details overview of machine learning. They analyzed the best and recent research solid state materials science. They also description the topic for the different machine learning techniques in the field of constant things. They used this approach to predict of crystal structure. They reviewed and presented active learning technique and also explain surrogate based optimization. They explain that these techniques can be useful to increase the rational design procedure [13].

In 2020 Yura Malitsky et al proposed "Adaptive Gradient Descent without Descent". They explained and presented by using a simple example that two guidelines are enough to mechanize gradient descent. These guideline are first one is there is no need to increase the step size too fast. Second one is there is no need overstep the local curvature. With the help of these guidelines they get a adaptive technique to the native geometry. By using convergence there are assurances for smoothness and depending only on the neighborhood for a solution [15].

In 2020 Nam D. et al proposed **"**Implicit Stochastic Gradient Descent Method for Cross-Domain Recommendation System**".** They used old system of recommendation and applied the factorization and collaborative filtering technique. They used this technique with single domains. The proposed approach has a limited cold start problem. They removed data sparsity and remove the limitation. They focused and discovered several features from various domains. They describe and explained the relationships between domains. They considered and applied proposed approach for multiple domains [16].

3. PROBLEM WITH GRADIENT DESCENT

Gradient descent techniques has very good convergence, but this approach has a few problems which are given below [5,6]

3.1 Rate of learning

This is difficult parameter and we cannot select the value of learning rate easily. Small value rate of learning create very slow convergence and difficult to handle. Large value of rate of rate can delay convergence. This will create the cause of the loss function and oscillate round the minimum.

3.2 A lot of local minima

There are so many Local Minima. If there are several local minima presented, in this situation there is no guarantee to found global minimum. The procedure will take large number of iterations. If the data used for the technique is sparse and features have large number of different frequencies, it is very difficult to update all of them because we are using same extent. To it is rarely that the performance for large update occurring.

3.3 Reducing non-convex error

There are several functions common are used in neural networks. These functions are avoiding and getting trapped for local minima. These function help when there are several points has one dimension slopes up and one dimension slope down. Due to these lumber points plateau of the same error are usually surrounded and which makes it notoriously hard.

4. APPLY TO RIGHT GRADIENT DESCENT

It is very difficult to select right gradient descent technique for a give problem. It is a best technique to make definite gradient descent tracks properly. We have to use plotting for the cost function and try to find the number of optimization runs. For the plotting we place the number of repetitions on the axis of x and the charge of the function of cost on the axis of y.

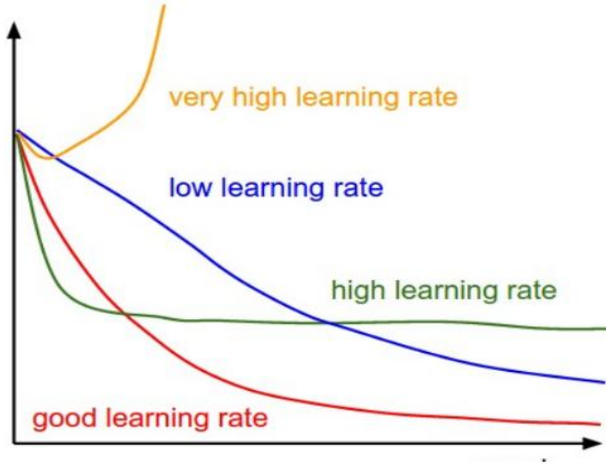

Fig 2: Different value of learning rate

This plotting of graph helps to understand the charge function of cost. After completing each iteration of gradient descent we need to check it. This value delivers a technique to simply advert how appropriate rate of learning. We can use different values and plot them all collected. The plotted image appearances such a plot, while the image illustrates the change among good and bad rates of learning

5. STEPS TO IMPLEMENT GRADIENT DESCENT

Some of the important steps and tricks essential to use before implementing gradient descent technique for best machine learning [8, 9, 10]

1. Plotting of Cost versus Time: We need to plot the cost values and time. These values are calculated by the algorithm. In each and every iteration we have to collect these values. The expectations for good performance gradient descent run with decrease in cost.
2. Rate of Learning: The values for rate of learning rate are always small. Small real values such as 0.1, 0.001 or 0.0001 are used for rate of learning. Sometimes different values are also used for difficult and see which mechanism best.
3. Inputs are rescaled: The techniques need to rescale the input. When the technique reached with minima the shape of the function of cost is not skewed and biased. We can achieved by rescaling values of all input variables to the similar series, such as [0, 1] or [-1, 1].
4. Limited Passes: Stochastic gradient technique is frequently not necessity more than 10 passes. The training dataset to converge on good enough coefficients.
5. Mean cost need to plot:- The values updated for each training example resulted with a noisy plot. When using stochastic gradient techniques there are average over ten, hundred, or thousand updates. This gives a better idea of the learning trend for the technique.

6. TYPES OF GRADIENT DESCENT

Three most commonly types of Gradient are [11,12,14]:
1. Gradient based Batch
2. Stochastic Gradient
3. Gradient based on Mini-Batch

6.1 Batch Gradient Technique

The first and most basic kind of gradient technique. In this technique we used the whole set of data and compute gradient with function of cost. We essential calculate the gradient for the complete dataset to perform only in single update. Batch gradient technique is very slow and is obstinate for datasets that don't fit in memory.

6.2 Stochastic Gradient

Batch Gradient is very slow. Due to this problem we essential want fast computation, so most of the people prefer stochastic gradient technique. The first step used in this technique randomizes the whole training data set. In the nest step we update every parameter. We use only one training example. In every repetition we must to compute the gradient of cost function. As we used single training example so technique is faster for larger data set. This technique might be possible not achieve accuracy, but the computations of results are faster.

6.3 Gradient based on Mini-Batch

Gradient based on Mini batch is the very satisfactory and largely used technique which types correct and faster results. We used a batch of training examples. In this technique we create 'm' training examples in place of the complete data set. In every repetition we use to set value of 'm' training examples called batch to compute the gradient. We need to calculate function of cost. Most normally mini sizes of batches range between fifty and two hundred fifty but it can be possible for different values for different applications.

7 COMPARATIVE BASED PROS AND CONS

Table 1 comparative based pros and cons

Gradient Descent technique	pros	cons
Stochastic	Easy and suitable in small memory. Computation faster Efficient for large data sample	Steps used towards minima Frequent updates it is noisy. Noise create large to wait. Due to Frequent updates expensive computationally
Batch	Less noisy. Steps are stable convergence. Computationally efficient Resources are not used for single sample	Additional memory required and needed. It can take long to process for large database. Approximate gradients
Mini Batch	Easy fit in memory. Computationally efficient. Stable error go and convergence.	Sometimes a constant error can lead to local minima. Complete training data set can be very large to proceed. Additional storage might be required

8. CONCLUSION

In this paper we proposed detailed study of different Gradient techniques like Batch Gradient Descent technique, Stochastic Gradient Descent technique and Mini batch Gradient Descent technique. These can be used when the dataset is large or small. A batch Gradient technique converges directly to minima. We use only one example at a time, we cannot implement for vectorized data. It can slow down the computations. Sometimes we need to use a mixture of Batch Gradient techniques is used. Average cost over the epochs in mini-batch gradient descent fluctuates because averaging a small number of examples at a time.

8. REFERENCES

[1] Yiming Ying et al "Online gradient descent learning algorithm" Department of Computer Science, University College London Gower Street, London, 2014 WC1E 6BT, England, UK.

[2] Diederik P. Kingma et al "Adam: A Method For Stochastic Optimization" arXiv:1412.6980v9 30 Jan 2017.

[3] Marcin Andrychowicz et al "Learning to learn by gradient descent by gradient descent". Western Norway Research Institute, Box 163, NO-6851 Sogndal, Norway

[4] Stephan Mandt and Matthew D. Hoffman Stochastic Gradient Descent as Approximate Bayesian Inference Journal of Machine Learning Research 18 (2017) 1-35 Submitted 4/17; Revised 10/17.

[5] Sebastian Ruder "An overview of gradient descent optimization algorithms". Insight Centre for Data Analytics, NUI Galway Aylien Ltd., Dublinruder

[6] Shuang Song et al "Stochastic gradient descent with differentially private updates" Dept. of Computer Science and Engineering University of California, San Diego La Jolla, CA USA.

[7] Loucas Pillaud-Vivien et al "Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes" INRIA – Ecole Normale Supérieure PSL Research University

[8] Leon Bottou et al "Optimization Methods for Large-Scale Machine Learning" 2018 Society for Industrial and Applied Mathematics Vol. 60, No. 2, pp. 223–311.

[9] Prateek Jain et al "Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification" Journal of Machine Learning Research 18 (2018) .

[10] Nan Cui "Applying Gradient Descent in Convolutional Neural Networks" CMVIT IOP Publishing IOP Conf. Series: Journal of Physics: Conf. Series 1004 (2018) .

[11] E. M. Dogo et al " A Comparative Analysis of Gradient Descent-Based Optimization Algorithms on Convolutional Neural Networks" 978-1-5386-7709 2018 IEEE.

[12] Dokkyun Yi et al "An Enhanced Optimization Scheme Based on Gradient Descent Methods for Machine Learning" Daegu University, Kyungsan 38453, Korea 8 June 2019.

[13] Jonathan Schmidt et al "Recent advances and applications of machine learning in solid state materials" science 26 February 2019 Accepted: 17 July 2019.

[14] Simon Shaolei et al "Gradient Descent for Non-convex Problems in Modern Machine Learning" APRIL 2019 School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213.

[15] Yura Malitsky Konstantin et al "Adaptive Gradient Descent without Descent" Proceedings of the 37 th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[16] Nam D. Vo et al "Implicit Stochastic Gradient Descent Method for Cross-Domain Recommendation System Sensors" 2020, Western Norway Research Institute, Box 163, NO-6851 Sogndal, Norway.