

# On Gradient Descent and Coordinate Descent Methods and Its Variants

Sajjadul Bari, Md. Rajib Arefin, and Sohana Jahan

**Abstract**—This research is focused on Unconstrained Optimization problems. Among a number of methods that can be used to solve Unconstrained Optimization problems we have worked on Gradient and Coordinate Descent methods. Step size plays an important role for optimization. Here we have performed numerical experiment with Gradient and Coordinate Descent method for several step size choices. Comparison between different variants of Gradient and Coordinate Descent methods and their efficiency are demonstrated by implementing in loss functions minimization problem.

**Index Terms**—Convex function, Coordinate descent, Differentiable function, Gradient descent, Lipschitz constant, L-smooth function, Unconstrained optimization.

## I. INTRODUCTION

UNCONSTRAINED optimization [12], [15] problem minimizes an objective function that depends on real variables with no restrictions on their values. Mathematically, if  $x \in \mathbb{R}^n$  is a real vector with  $n \geq 1$  components and if  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is a smooth function. Then, the unconstrained optimization problem is of the form

$$(P) \min_x f(x) \\ \text{s.t. } x \in \mathbb{R}^n$$

Thus we want to find an optimal decision, that is  $x^* \in \mathbb{R}^n$  such that

$$f(x^*) \leq f(x), \quad \forall x \in \mathbb{R}^n.$$

Unconstrained optimization have been used in applications for many years, and their popularity continues to grow because of their usefulness in data analysis, machine learning, and other areas of current interest. Unconstrained optimization problems may arise directly in many applications or they may arise from reformulations of constrained optimization problems. Constraints of an optimization problem can be replaced in the objective function with penalized terms and the constrained optimization problem can be solved as an unconstrained problem. In this paper we have worked on iterative techniques for solving unconstrained optimization problem specifically on Gradient descent method and Coordinate Descent method.

Sajjadul Bari is with Department of Mathematics at American International University- Bangladesh as a Lecturer (e-mail: sajjadul@aiub.edu).

Md. Rajib Arefin is with Department of Mathematics at University of Dhaka as Assistant Professor. He is now on study leave for pursuing PhD at Kyushu University, Japan. (e-mail: arefin.math@gmail.com).

Sohana Jahan is with Department of Mathematics at University of Dhaka as Associate Professor. (e-mail: sjahan.mat@du.ac.bd).

## A. Main Features of Algorithms

General structure of an iterative algorithm for solving unconstrained minimization problem is as follows

- Choose a starting point  $x_0$ .
- Beginning at  $x_0$ , generate a sequence of iterates  $\{x_k\}_{k=0}^{\infty}$  with non-increasing function (f) value until a solution point with sufficient accuracy is found or until no further progress can be made.

To generate the next iterate  $x_{k+1}$ , the algorithm uses information about the function at  $x_k$  and possibly earlier iterates.

Both the Gradient Descent method and Coordinate descent method follow the above two steps in minimizing a function.

- Step Length: A suitable step-length can help the initial guess to reach the goal in the fastest way. If the step-length is too long it may exceed the target and on the contrary, if it is too short the convergence will be slow.
- Descent Direction: At each iteration a descent direction has to determine. This direction is opposite to the direction of gradient of the function at the current point,

Here we will concentrate on the discussion of different aspects of Gradient Descent and Coordinate Descent method which will assist us to find a right descent direction.

The remaining part of this paper is organized as follows: In the next section we have discussed choice of step length and the respective algorithms of for Gradient descent method. The following section includes the discussion of co-ordinate descent method. Choice of co-ordinates to update at each iteration is also discussed briefly. A comparison between GD and CD for different choice of stepsize is shown numerically in section IV. In section V we have implemented the idea of GD and CD in loss function minimization problem. We have concluded our results in section VI.

## II. GD: GRADIENT DESCENT

The basic Gradient Descent Method [3], [10], [11], [4] is based on fixed step size. A variant of GD includes choice of different step size so that the algorithm performs efficiently. Moreover, step length can be chosen with backtracking armijo condition to get a better approximation. In this section we will discuss each of these variants of GD.

The following algorithms [3], [5], [4], [14] is on Gradient Descent Method of different variants. step size.

Algorithm 1.(Gradient Descent with Fixed Step Size)  
The algorithm is initialized with a guess  $x_0$ , a maximum











TABLE XII: Some iterations of CD on  $f_2(x, y)$  with backtracking armijo condition

Iteration No.	$ f_2(x_{exact}) - f_2(x) $
1	21.192545
2	10.974881
⋮	⋮
47	$1.66 \times 10^{-8}$
48	$8.11 \times 10^{-9}$
49	$3.52 \times 10^{-9}$
50	$3.34 \times 10^{-9}$

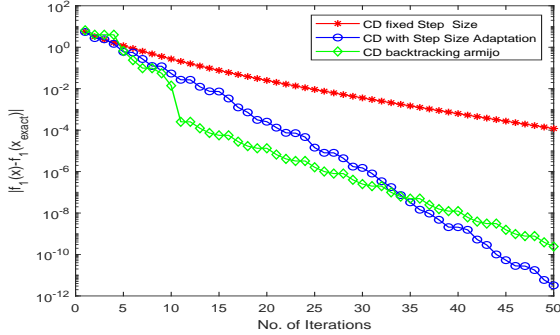
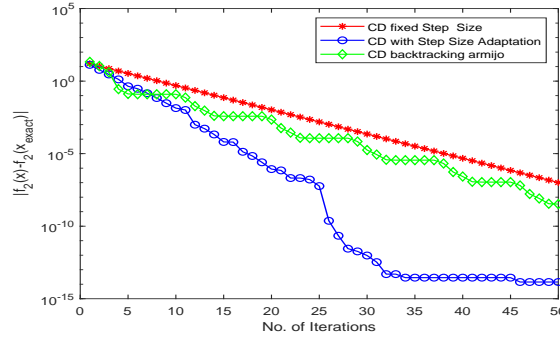
(a) on  $f_1(x, y)$ (b) on  $f_2(x, y)$ Fig. 14: Comparison on CD with different step sizes for  $f_1(x, y)$  in (a) and  $f_2(x, y)$  in (b)

Figure 14 (a) and (b) gives the comparison of three cases of CD on both  $f_1(x, y)$  and  $f_2(x, y)$  respectively in the same frame which implies that step size adaptation technique gives better approximations for coordinate descent method but if we stop before 35<sup>th</sup> iteration (figure (a)) then backtracking armijo technique gives more accuracy than step size adaptation. For a better accuracy one can insert the stopping criterion.

$$|f(x) - f(x_{exact})| < \epsilon; \text{ suppose } \epsilon = 10^{-05}$$

The more the  $\epsilon$  tends to zero, the more accuracy will be obtained. After a certain number of iterations, step size adaptation technique will perform better than backtracking armijo.

Comment :

The resulting comparison on GD and CD may vary considering different choices of functions.

## V. LOSS FUNCTION MINIMIZATION

We are going to define loss function [16], [5], [17] as

$$f(x) = \frac{1}{n} \sum_{i=1}^n (a_i^T x - b_i)^2$$

Where  $a_i$  is a row vector,  $x$  and  $b_i$  are column vectors.

This is the least-squares loss function that gives rise to the ordinary least squares regression model. The loss function is obviously convex function. Minimizing an arbitrary function is, in general, very difficult, but if the objective function to be minimized is convex then things become considerably simpler. The key advantage of dealing with convex function is that a local optima is also a global optima.

We will concentrate on numerical experiment with  $L_2$ -regularized least square problem using gradient descent and coordinate descent algorithm. We consider,

$$f(x) = \frac{1}{n} \sum_{i=1}^n (a_i^T x - b_i)^2 + \frac{\lambda}{2} \|x\|^2$$

The main goal is to predict  $x$  that minimizes the loss function

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$$

The function  $f_1, f_2, f_3, \dots, f_n$  are assumed to be  $L$ -smooth. Clearly,  $(a_i^T x - b_i)^2$  is convex, therefore,  $f(x)$  is strongly convex<sup>1</sup> with  $\lambda$ .

We can estimate the Lipschitz constant  $L_i$  for the function  $f_i$  as  $(2\|a_i\|^2 + \lambda)$ . Thus Lipschitz constant for  $f(x)$  would be  $\max_{1 \leq i \leq n} \{L_i\}$ . In this case, a training set of 50 examples are being considered. Each example comprises 30 features. That is  $n = 50$  and  $d = 30$ . The entries of  $a_i$  are taken as random integers from 1 to 10. The constant  $\lambda$  is considered as  $\frac{1}{n}$ .

We have applied both GD and CD and run the algorithms for 100 iterations.

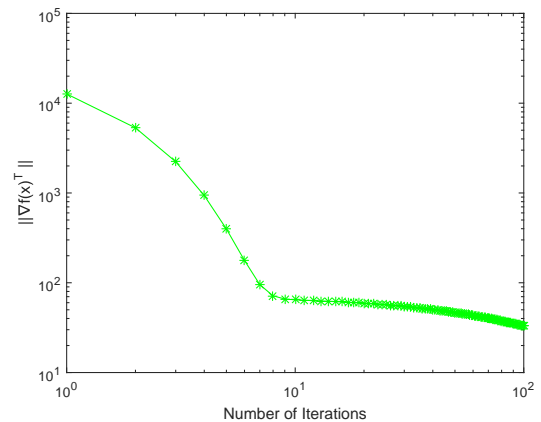


Fig. 15: Loss function minimization with GD.

<sup>1</sup>A convex function  $f$  is strongly convex if and only if, there exist a constant  $\mu > 0$  such that the function  $f(x) + \frac{\mu}{2} \|x\|^2$  is convex.

TABLE XIII: Some observation of GD for loss function minimization

Iteration No.	$\ \nabla f(x)\ $
1	12678.71
2	5314.989
⋮	⋮
97	33.5855
98	33.37241
99	33.16144
100	32.95256

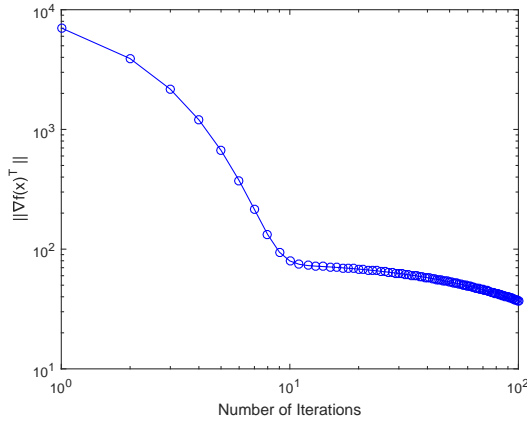


Fig. 16: Loss function minimization with CD.

TABLE XIV: Some observation of CD for loss function minimization

Iteration No.	$\ \nabla f(x)\ $
1	7038.337
2	3906.364
⋮	⋮
97	37.80994
98	37.55346
99	37.29952
100	37.04807

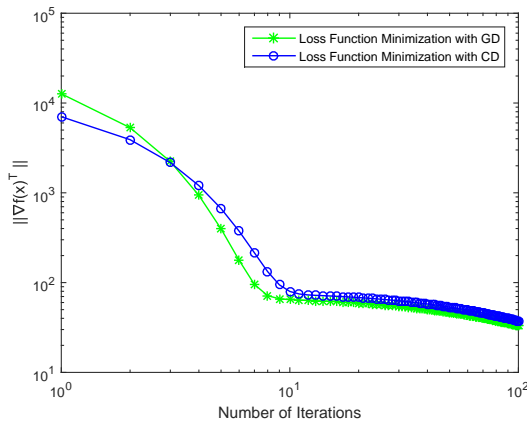


Fig. 17: GD vs CD for Loss function minimization.

Figure 17 represents the performance of GD and CD in loss function minimization for 100 iterations. It is clearly seen that GD works better than CD. But for a large number of iterations the performance of two methods are quite similar.

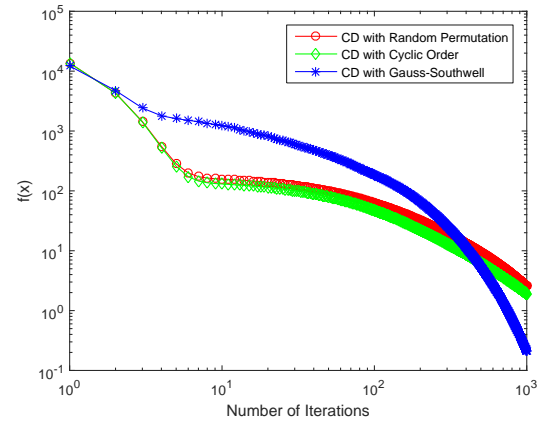


Fig. 18: GD vs CD for Loss function minimization.

We have applied CD on loss function minimization with different choice of selecting coordinates for the update. From figure 18 it can be concluded that though at the initial stage Gauss-Southwell was slow in convergence but after a certain iteration it works better then other two techniques.

## VI. CONCLUSION

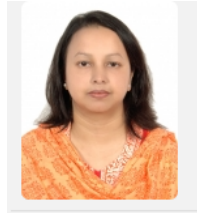
In this paper we have worked on unconstrained optimization problem. Specially we are focused on performance of iterative techniques Gradient Descent (GD) method and Coordinate Descent (CD) method in solving the unconstrained optimization problem. We have applied both GD and CD for different choice of step size and check their performance in minimizing a problem. Numerical results show that for both GD and CD method step size adaptation technique converges faster until a certain number of iterations. And also we may conclude that this comparison may vary for different choices of function and step size. A comparison between these two for solving loss function minimization is also demonstrated which shows that for loss function minimization problem GD works better than CD in general. In case of CD, Gauss-Southwell technique can be used for the updates of coordinates at different iteration for getting faster convergence.

## REFERENCES

- [1] J. Désidéri. "Multiple-gradient descent algorithm (MGDA) for multiobjective optimization," *Comptes Rendus Mathématique*, vol. 350, no. 5, pp. 313-318, Mar. 2012. DOI: <https://doi.org/10.1016/j.crma.2012.03.014>
- [2] M. Blondel *et al* "Block coordinate descent algorithms for large-scale sparse multiclass classification," *Machine Learning*, vol. 93, no. 1, pp. 31-52, May 2013. DOI: <https://doi.org/10.1007/s10994-013-5367-2>
- [3] S. Ruder, "An overview of gradient descent optimization algorithms," Jun. 2017. <https://arxiv.org/abs/1609.04747v2>
- [4] S. Wright, and J. Nocedal. "Numerical optimization," *Springer Science*, vol. 35, no. 7, pp. 35 (1999): 67-68.
- [5] M. R. Arefin, and M. Asadujjaman., "Minimizing Average of Loss Functions Using Gradient Descent and Stochastic Gradient Descent," *The Dhaka University Journal of Science*, Vol. 64, no. 2, pp. 141-145, Jul. 2016.



- [6] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3-34, Mar. 2015.
- [7] P. Richtárik, and M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming*, vol. 144, no.1-2, pp. 1-38, Apr. 2014.
- [8] A. Beck, and L. Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2037-2060, Oct. 2013.
- [9] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341-362, Apr. 2012.
- [10] Y. Nesterov, "Introductory Lectures on Convex Optimization, A Basic Course," Springer, 2003.
- [11] S. Boyd *et al.*, "Unconstrained minimization," in *Convex Optimization*, 1st ed., Cambridge University Press, 2004, ch. 9, pp. 457-475. DOI: <https://doi.org/10.1017/CBO9780511804441>
- [12] J. Nocedal, "Theory of algorithms for unconstrained optimization," *Acta Numerica*, vol. 1, pp. 199-242, Jan. 1992. DOI: <https://doi.org/10.1017/S0962492900002270>
- [13] X. Wang, "Method of steepest descent and its applications," *IEEE Microwave and Wireless Components Letters*, vol. 12, pp. 24-26, Nov. 2008.
- [14] D.P. Mandic, "A generalized normalized gradient descent algorithm," *IEEE signal processing letters*, vol. 11, no. 2, pp. 115-118, Feb. 2004. DOI: <https://doi.org/10.1109/LSP.2003.821649>
- [15] Astolfi, A. "Optimization, An Introduction." Estados Unidos, Octubre del (2005).
- [16] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific Journal of mathematics*, vol. 16, no. 1, pp. 1-3, Jan. 1966. DOI: <https://doi.org/10.2140/pjm.1966.16.1>
- [17] D.P. Bertsekas, "Unconstrained Optimization," in *Nonlinear Programming*, 2nd ed. Belmont, Massachusetts, Athena scientific, 1999, ch. 1, sec. 1-8, pp. 4-160.



**Sohana Jahan** received the PhD in Operational Research from School of Mathematics, University of Southampton, UK in 2017, obtaining the Commonwealth Scholarship for that PhD program. She got A. F. Mujibur Rahman foundation gold medal for excellence in Pure Mathematics. She also acquired Mitra-Yusuf trust scholarship for the best result during third year of undergraduation.

Currently she is working as an Associate Professor at Department of mathematics, University of Dhaka. Before that she joined at University of Dhaka as a Lecturer in 2010 and promoted as an Assistant Professor in 2017. She also worked as a Lecturer at department of Mathematics, BUET in 2010.



**Sajjadul Bari** completed his M.S and B.S(Hons) degree from University of Dhaka at Department of Mathematics. He became a life member in Bangladesh Mathematical Society in 2020. His research works focus on Operations Research. He also grew his interest on some topics with collaboration of other fields of study. In addition to all these, he is also studying books and articles related to Graph Theory to pursue his higher education later.

He is currently working as a Lecturer in Mathematics at American International University-Bangladesh. He has been working here since September 10, 2018. Prior to that, he was employed with Daffodil International University as a Lecturer. He participated several workshops, webinars and trainings to gather professional skills.



**Md. Rajib Arefin** completed his MSc in Financial Operational Research (FOR) (with Distinction), from University of Edinburgh, UK. He has received Dean's award from the faculty of Science, University of Dhaka for extraordinary performance in BSc (Hons.) examination. He also has been honored by A F Mujibur Rahman gold medal (twice) for securing first position in BSc and MS examinations. Besides these, he acquired University Grants Commission (UGC) of Bangladesh scholarship for excellence in Mathematics in BSc examination.

His Research interests lie in the field of Operations Research. He has also interest in applying modern optimization techniques in financial models. His another interest is evolving into the numerical simulations of Stochastic Differential Equations (SDEs).

He is currently working as an Assistant Professor (on study leave for pursuing PhD at Kyushu University, Japan) in Department of Mathematics at University of Dhaka. Before that, he had been working as a Lecturer in the same place since 2015. He had been teaching several undergraduate courses such as, Calculus, Differential equations, Mathematical Methods, Linear Algebra, Introduction to Mathematical Finance for the last couple of years. He taught an MSc level course: Operations Research as well. Besides all these, he supervised several undergraduate project students and also completed supervision of a few M.S (Thesis) students.