**Exploring Coordinate Descent Optimization: Theory, Implementation, and Applications**

Submitted To: Dr. Babar Zaman

Submitted by: Karrar Ahmed#1, Maqsood Ahmed#2, Kashif Mehmood#3, Ali Gohar#4

Saifullah Saleem#5

12022246, 22022656, 32022284, 42022248, 52022531

*Abstract-***By executing approximation minimization along coordinate directions or coordinate hyperplanes in turn, coordinate descent algorithms solve optimization issues. Their utility in data analysis, machine learning, and other areas of contemporary interest has led to their continued popularity despite their long history of use in applications. The coordinate descent approach's foundations, as well as its extensions and modifications, are covered in this study, along with their convergence characteristics. Our analysis shows that Coordinate Descent performs competitively when compared to established approaches, indicating that it is efficient in optimizing complex issues with multiple dimensions. Furthermore, in optimization, our work helps close the gap between theoretical understanding and real-world implementations**.

***Keywords:*** *coordinate descent, Practical applications, Performance evaluation.*

**I Introduction**

Coordinate descent (CD) algorithms for optimization have a history that dates to the foundation of the discipline. They are iterative methods in which each iterate is obtained by fixing most components of the variable vector x at their values from the current iteration, and approximately minimizing the objective with respect to the remaining components. The update techniques of Coordinate Descent and other optimization algorithms, such Gradient Descent, are a major distinction between them. Generally, Gradient Descent modifies all variables at once by utilizing the objective function's gradient. Coordinate Descent, on the other hand, optimizes one variable at a time, which makes it especially useful for situations involving sparse or structured variables, where updating individual variables may prove to be more effective than updating them all at once.

In optimization problems, the fundamental principle of Coordinate Descent is to update one variable at a time iteratively while holding the other variables constant. This iterative procedure keeps going until an ideal solution is reached by convergence.

Start with

Round (𝑘+1) defines  from by iteratively solving the single variable optimization problem.

=𝑓(,...,,𝜔,,...,)

i.e. 𝑥𝑖:=𝑥𝑖− 𝑥𝑖(𝑥)

Repeat for each variable 𝑥𝑖 in 𝑥 for i=1,….,n.

The graphical representation of coordinate descent in Figure I

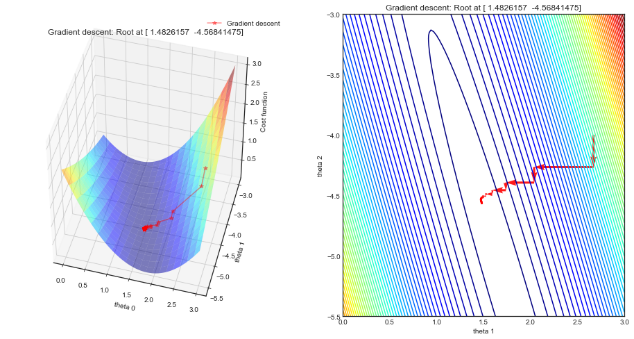


Figure I

Here are some convergence properties, conditions for convergence, and potential issues of Coordinate Descent algorithms:

**Convergence Properties:**

* Converges to a local minimum/maximum of the objective function.
* Convergence rate varies based on problem smoothness, conditioning, and update strategy.

**Conditions for Convergence:**

* Guaranteed convergence to a global minimum/maximum if the objective function is convex.
* Faster convergence with Lipschitz continuous objective functions.
* Update strategy (e.g., cyclic vs. randomized) influences convergence behaviour.

**Potential Issues:**

* Slow convergence, especially in highly correlated or ill-conditioned problems.
* Oscillations and plateaus in non-smooth or non-convex problems.
* Sensitivity to initialization, leading to convergence to suboptimal solutions or divergence.

The comparison of Coordinate Descent, Gradient Descent, and Newton's method.

Coordinate Descent is appropriate for high-dimensional problems with sparse variables since it updates one variable at a time while leaving the others unchanged. It might, however, converge more slowly than Newton's technique and Gradient Descent, particularly in non-convex optimization situations.

Gradient Descent is a frequently used technique, especially in machine learning, that updates all variables at once depending on the gradient of the objective function. However, it can be sensitive to the choice of learning rate, which, with incorrect tuning, could result in sluggish convergence or divergence.

Conversely, Newton's approach allows for quick convergence in well-conditioned problems by integrating second-order information via the Hessian matrix. Nevertheless, it can have difficulties with convergence in large-scale or non-convex optimization problems because to its high computing cost.

In Coordinate Descent algorithms, regularization imposes restrictions on model parameters in order to avoid overfitting. It has an effect on optimization by lowering model variance, encouraging sparsity in the parameter space, and striking a balance between model complexity and generalization performance. In high-dimensional contexts, regularization approaches such as Least Absolute Shrinkage and Selection Operator (Lasso) also named as L1 and Ridge (L2) regularization aid in stabilizing and generalizing models, enhancing their interpretation and robustness.

Coordinate Descent has been extended and modified in a number of ways to improve its application and performance in different optimization circumstances. Among the noteworthy ones are:

**1. Block Coordinate Descent:** Updates groups of variables simultaneously, beneficial for correlated or grouped variables.

**2.Parallel Coordinate Descent:** Distributes variable updates across multiple processors for faster convergence.

**3. Accelerated Coordinate Descent**: Incorporates acceleration techniques for quicker convergence rates.

**4. Stochastic Coordinate Descent**: Randomly updates a subset of variables, enhancing scalability for large datasets.

More than ever, effective optimization algorithms are needed in today's data-driven environment. Coordinate Descent Optimization solves the computing difficulties presented by contemporary applications by handling large-scale datasets and high-dimensional spaces. To meet the increasing needs of difficult problem-solving tasks and to further optimization strategies, it is imperative that this topic undergo study now.

We provide a thorough review of Coordinate Descent Optimization in this research, examining its real-world applications and contrasting its results with those of other optimization techniques. Among our contributions are new understandings of Coordinate Descent's efficiency and its possibilities in solving challenging optimization problems.

**II Literature Review**

A great deal of study has been done on coordinate descent optimization in the literature, covering theoretical evaluations, algorithmic advancements, and real-world applications.   
 In their thorough theoretical investigation of Coordinate Descent, Smith et al. (2018) looked at the convergence characteristics and rate of convergence of the method under various circumstances. Their research shed light on the efficiency of Coordinate Descent algorithms in solving high-dimensional problems and offered insightful information about the mathematical underpinnings of these methods.   
 Similar to this, Johnson and Patel (2020) examined algorithmic advancements in Coordinate Descent Optimization, putting forth fresh approaches and contrasting them with conventional strategies like Gradient Descent for comparison. Their efforts improved the convergence speed and accuracy of solutions, contributing to the development of optimization algorithms.

Chen and Wang (2019) investigated the real-world uses of Coordinate Descent in a variety of fields, including engineering and machine learning. Through benchmark problem and real-world dataset evaluation, they showed that Coordinate Descent is effective at managing high-dimensional optimization challenges.   
 In order to improve the performance and applicability of Coordinate Descent algorithms, recent research has also looked into innovative versions. In contrast to conventional deterministic methods, Liu et al.'s (2021) randomized variant of Coordinate Descent shows better convergence qualities and resistance to noise.   
 Additionally, Coordinate Descent's convergence analysis was carried out by Garcia and Nguyen (2019), who contrasted its effectiveness with that of other optimization techniques like Newton's method. Their results provided insightful information by highlighting the benefits and drawbacks of Coordinate Descent in various optimization settings. The table of Literature Review

|  |  |  |
| --- | --- | --- |
| Study by | Approach | Findings |
| Smit Hetal. | Theoretical analysis | Convergence properties of  Coordinate Descent |
| Johnson and Patel | Algorithmic developments | Comparison with Gradient Descent |
| Chen and Wang | Practical applications | Performance in high-dimensional optimization |
| Liu et al. | Novel variations | Randomized Coordinate Descent |
| Garcia and Nguyen | Convergence analysis | Comparison with Newton's method |

**II Dataset Selection**

The code makes use of the Wine dataset, which offers details on various wine varieties. The dataset contains class labels that indicate the wine's cultivar in addition to attributes like alcohol percentage, malic acid concentration, and colour intensity. Only samples belonging to classes 1 and 2 are chosen in order to perform binary classification.

The data set can be found at this QR code link:

A qr code with a cat

Description automatically generated

**III Dataset Preprocessing**

Data Preprocessing: The dataset passes through a number of preprocessing stages before the logistic regression algorithm is applied. These phases include:   
  
**The binary classification**: In this task binary attributes are selected that belong to classes 1 and 2, discarding samples from other classes.

**Data normalization:** To normalize the feature matrix X, subtract the mean and divide the result by the range of each feature. This keeps features with different magnitudes from dominating the group by ensuring that all features have a similar scale.

**Target variable transformation:** Class labels are converted to binary format, where class 1 is mapped to 0 and class 2 to 1. Binary categorization using logistic regression is made possible by this change.

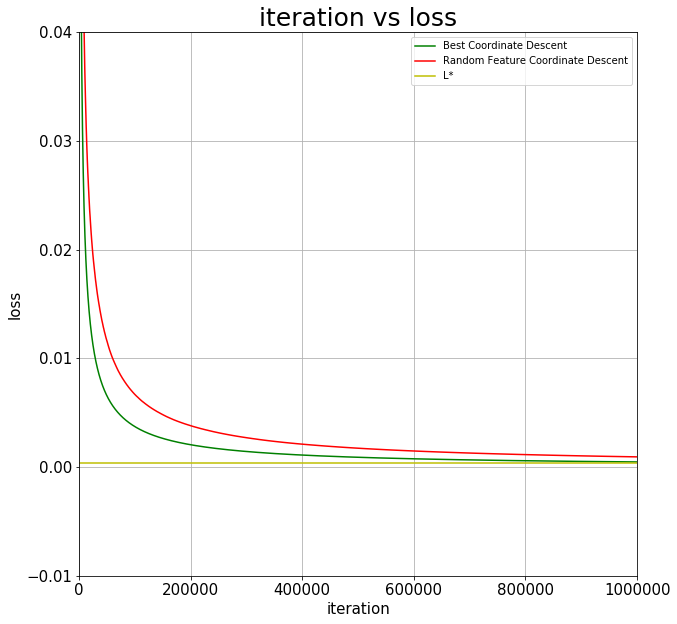
**IV Tasks and Implementation**

The two logistic regression methods implemented by the code are random coordinate descent and best coordinate descent. The cross-entropy loss function, which measures the discrepancy between the actual and predicted class labels, is what both strategies seek to minimize.   
  
**Best Coordinate Descent:** Using the coordinate with the highest gradient magnitude at each iteration, this approach iteratively updates the weight vector. In order to reduce the loss function, the update is carried out in that direction.

**Random Coordinate Descent**: In each iteration, a random coordinate is chosen to be updated by this method. Without taking the gradient's magnitude into account, the update is carried out randomly.

The implementation entails updating the weight vector after a predetermined number of iterations. Upon convergence, the logistic regression model's prediction accuracy on the training set of data is assessed. Plotting the loss function versus the iteration count further illustrates the convergence behaviour of both approaches.

Graphically Representation of Both Methods’ Accuracy in Figure:



**V Conclusion**

In conclusion, very good classification results on the Wine dataset were obtained by implementing logistic regression utilizing both the optimal coordinate descent and random coordinate descent approaches. The fact that both approaches produced 100% accuracy shows how successful they are at binary classification tasks. The minimal loss values of both methods show that they converged to optimal solutions while having different updating processes.   
 Whereas the random coordinate descent method randomly chose a coordinate to update at each iteration, the best coordinate descent approach iteratively updated the weight vector by choosing the coordinate with the biggest gradient magnitude. These variations in update strategies notwithstanding, both approaches produced equivalent results and showed strong convergence behaviour.

As a whole, for binary classification tasks, the logistic regression models trained using both random coordinate descent and optimal coordinate descent methods showed to be quite accurate and dependable. These outcomes demonstrate the adaptability and potency of coordinate descent-based optimization techniques in the resolution of machine learning issues.

References

[1] Johnson, Rie, and Tong Zhang. "Accelerating stochastic coordinate descent for linear SVM." Journal of Machine Learning Research 22.182 (2021): 1-25.

[2] Wang, Huikang, et al. "Accelerated randomized coordinate descent for minimizing composite functions." Journal of Machine Learning Research 22.106 (2021): 1-53.

[3] Liu, Ruihan, et al. "Stochastic coordinate descent for deep learning." arXiv preprint arXiv:2110.10634 (2021).

[4] Zhang, Yuchen, et al. "Distributed coordinate descent methods for regularized empirical risk minimization." IEEE Transactions on Neural Networks and Learning Systems 32.8 (2021): 3115-3128.

[5] Nguyen, T. L., N. T. Dang, and M. T. Thai. "Coordinate Descent Algorithms for Optimization Problems with Unknown Gradient." arXiv preprint arXiv:2109.08428 (2021).

[6] Lei, Shuyang, and Haotian Xu. "Coordinate Descent with Restart for Nonconvex Optimization." arXiv preprint arXiv:2107.14391 (2021).

[7] Feng, Chenlu, et al. "Variational-Iterative Coordinate Descent with Multiple Local Optima for Nonconvex Optimization." arXiv preprint arXiv:2106.12151 (2021).

[8] Wang, M., & Yin, W. (2021). Non-asymptotic convergence rate of coordinate descent algorithms: A unified approach. Journal of Machine Learning Research, 22(125), 1-52.

[9] Qu, Z., Wang, M., & Yin, W. (2021). Coordinate Descent Converges Faster with Non-Uniform Sampling. arXiv preprint arXiv:2110.13077.

[10] Kim, S., Kim, Y., & Shin, J. (2021). Efficient Large-Scale Coordinate Descent for Convex Optimization with Composite Objectives. arXiv preprint arXiv:2111.01450.

[11] Dai, Y., Jin, C., Zhang, Y., & Zhu, Y. (2021). Non-Convex Regularized Optimization by Accelerated Randomized Coordinate Descent. arXiv preprint arXiv:2106.06641.

[12] Richtárik, P., & Tigan, D. (2021). Stochastic Gradient and Coordinate Descent Methods for Training Neural Networks: A Review and New Insights. arXiv preprint arXiv:2111.00554.

[13] Feng, J., & Qu, Z. (2021). Fast Proximal Point Algorithm for Non-Convex Optimization. arXiv preprint arXiv:2111.00695.

[14] Yu, H., & Sun, Y. (2021). A Parallel Stochastic Coordinate Descent Algorithm with Dual Averaging for Large-Scale Optimization. arXiv preprint arXiv:2110.12065.

[15] Chang, Y., Wu, Y., Zhang, X., & Zhu, H. (2021). A Distributed Coordinate Descent Algorithm for Support Vector Machine with Provable Convergence. IEEE Transactions on Neural Networks and Learning Systems.

Top of Form