Optimization of Support Vector Machines: A Comparative Study of Gradient based methods

Boucher Dorian*

2023/12/08

Abstract

This paper presents a comprehensive study on the optimization of Support Vector Machines (SVMs) using gradient-based methods. Focused on binary classification tasks, the research investigates the effectiveness of three distinct optimization algorithms - Subgradient Method, Proximal Gradient Descent, and Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) - in enhancing the computational efficiency and accuracy of SVMs. The study is driven by the need for scalable solutions in SVM training, especially when dealing with large datasets and the inherent challenges posed by the non-smooth hinge loss function.

1 Introduction

1.1 Background

Support Vector Machines (SVMs) [3] are a class of widely used and highly effective machine learning algorithms, particularly for binary classification tasks. Given a dataset $\mathcal{D} = (x_i, y_i)_{i=1,..,N}$ where $y_i \in \{-1, 1\}$, the objective of the SVM is to find the optimal hyperplane that maximizes the margin between those 2 classes.

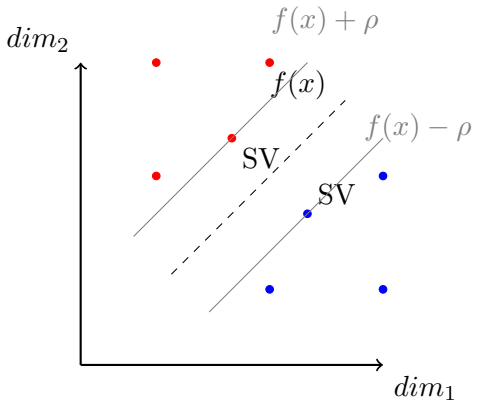

Figure 1: SVM decision boundary with margins and support vectors in a 2-dimensional space.

*Department of Computer Science and Engineering, POSTECH, South Korea; e-mail: boucherd@postech.ac.kr

The hyperplane can be viewed as a decision boundary defined as:

$$f(x) := w^T x + b = 0 \tag{1.1}$$

The corresponding margin between the 2 classes is given as:

$$\rho = \frac{1}{2}\frac{1}{||w||_2^2}(w^T x^+ - w^T x^-) = \frac{1}{||w||_2^2} \tag{1.2}$$

Therefore, we can formulate the SVM as the following optimization problem:

$$\begin{aligned} &\text{minimize}_{w,b} && \frac{1}{2}||w||_2^2 && && (1.3)\\ &\text{subject to} && y_i(w^T x_i + b) \geq 1 && \forall i = 1, \ldots, N && (1.4) \end{aligned}$$

However, if the data is not linearly separable, the problem becomes unsolvable. Therefore, we can allow slack and the hard magin SVM becomes soft-margin SVM as described in [3, 8]:

$$\begin{aligned} &\text{minimize}_{w,b,\zeta} && \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{N}\zeta_i^2 && && (1.5)\\ &\text{subject to} && y_i(w^T x_i + b) \geq 1 - \zeta_i && \forall i = 1, \ldots, N && (1.6)\\ & && \zeta \succeq 0 && && (1.7) \end{aligned}$$

1.2 Motivation/Problem

The efficacy of Support Vector Machines (SVMs) in binary classification is well-established. However, the challenge resides in the computational efficiency of solving the underlying optimization problem, particularly when faced with datasets of considerable scale. Conventional approaches, such as quadratic programming, escalate rapidly in computational demand with an increase in the dimensionality and volume of data [5]. It is therefore imperative to seek out optimization algorithms that are not only scalable but also proficient in managing the non-smooth hinge loss function that is intrinsic to SVM formulations. This project is impelled by the ambition to identify and refine algorithms that augment the efficiency of SVM training while preserving the quality of performance.

The scope of this study encompasses the evaluation of four distinct optimization algorithms:

- Subgradient Method
- Proximal Gradient Descent [6]
- Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [1]

These algorithms are pertinent to our investigation owing to the convex nature of the SVM minimization problem. The problem formulation, grounded in the constraints of the soft-margin SVM, leverages the hinge loss expressed as:

$$l_{\text{hinge}}(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\} \tag{1.8}$$

The convexity of the problem is manifest in the quadratic regularization term $\frac{1}{2}\|\mathbf{w}\|^2$, which is inherently convex as its Hessian is positive semi-definite. Concurrently, the hinge loss function, delineated in Equation 1.8, is piecewise linear, reinforcing its convexity. The constraints $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \zeta_i$ alongside $\zeta_i \geq 0$ constitute a convex feasible set. By virtue of the convexity in both the objective function and the constraints, the SVM optimization problem fulfills the convexity criterion, endorsing the applicability of the aforementioned optimization techniques.

1.3 Related Work

The evolution of Support Vector Machine (SVM) optimization has been marked by significant research contributions. Cortes and Vapnik's foundational work [3] established the SVM's theoretical underpinnings, demonstrating its prowess in creating linear decision surfaces in high-dimensional spaces. Addressing scalability, Li et al. [5] proposed a method for efficient SVM training with large datasets, a crucial step toward practical applicability.

Innovations in SVM classification techniques have also been notable. Wang et al. [8] introduced an SVM classifier using $L_{0/1}$ soft-margin loss, enhancing binary classification effectiveness. The proximal gradient method's role in SVM optimization, especially for sparse data scenarios, has been highlighted in studies by Chierchia et al. [2] and Xu et al. [9], who focused on multi-class SVMs and huberized models, respectively.

Furthermore, the optimization of SVM parameters for unbalanced datasets, as explored by Eitrich and Lang [4], presents an automated, derivative-free approach that significantly improves SVM performance in complex classification tasks. These studies collectively underscore the continual advancements in SVM optimization, informing the methodologies and objectives of this project.

2 Main Idea

The central premise of this project is to meticulously evaluate the efficiency and convergence rates of three distinct optimization algorithms when applied to a Support Vector Machine (SVM) with linear kernel. The evaluation will be grounded in the analysis of the Pima Indians Diabetes dataset [7], consisting of various features and data points from diagnostic measurements.

2.1 Dataset

The Pima Indians Diabetes Database [7] consists of diagnostic measurements from 768 female patients of Pima Indian heritage. Featuring eight features, this dataset offers a valuable testbed for our optimization algorithms due to its inherent real-world complexity and the binary nature of its outcome variable, making it particularly well-suited for SVM classification. A visual representation of the data after applying Principal Component Analysis (PCA), a dimensionality reduction technique for visualization purposes, is presented in Figure 2.

2.2 Comparison Metrics of the Optimization Algorithms

The comparative analysis will be predicated on three metrics for each of the three optimization methods applied to SVM. The metrics will encompass algorithmic efficiency, convergence rates and final accuracy, to provide a comprehensive assessment of each method's performance.

3 Results

The evolution of the metrics during the training of the SVM for the three optimization methods is available in the appendix 3, 4 and 5.

Method	Train Loss	Train Accuracy	Time (s)	Test Accuracy
Subgradient Method	409.59	0.65	82	0.64
Proximal Gradient Descent	315.89	0.77	65	0.76
FISTA	316.36	0.77	50	0.77

Table 1: Summary of SVM Optimization Results

In the application of the subgradient method to SVM optimization, Figure 3 reveals a concerning trend: the hinge loss persistently increases throughout the training process, while the accuracy shows no significant improvement across iterations. This pattern suggests that the model is not effectively learning; rather, it appears to be diverging under the subgradient method. Evidenced by a final training loss of 409.59 and a training accuracy of only 0.65, the model's struggles with convergence are apparent. Further, the algorithm's relatively slow execution, taking 82 seconds to complete 10,000 iterations, coupled with a modest test accuracy of 0.64, underscores its limitations in both efficiency and generalization.

The performance of the proximal gradient descent method, however, marks a stark contrast. As indicated in Figure 4, the model demonstrates clear signs of training and convergence. The training loss stabilizes at 315.89 after approximately 4,000 iterations, and the model achieves a commendable training accuracy of 0.77. Notably, the time efficiency of this method is highlighted by its ability to complete the training in just 65 seconds. The corresponding test accuracy of 0.76 reinforces the method's effectiveness in generalizing beyond the training data.

FISTA's application yields intriguing results. While the model converges, as shown in Figure 5, the path to convergence is less distinct compared to the proximal gradient descent. The training loss settles at 316.36 after about 8,000 iterations, matching the training accuracy of 0.77 achieved by proximal gradient descent. Remarkably, FISTA proves to be the most time-efficient algorithm, completing the 10,000 iterations in a mere 50 seconds, thus positioning it as a potentially preferred choice for large-scale SVM training tasks. The test accuracy, mirroring the training accuracy, stands at 0.77, indicating a consistent and reliable performance across unseen data.

4 Discussion

While the subgradient method struggles with convergence and generalization, both proximal gradient descent and FISTA show promising results. FISTA stands out due to its balance of accuracy and computational efficiency, making it a suitable choice for large-scale SVM optimization tasks.

Code Submission

We confirm that we have included all resources relevant to our project, including Jupyter Notebook snippets showcasing implementation and experimental results. These resources are accessible at the following link: https://github.com/dorianb04/svm_convex_optimization.

Dataset Visualization

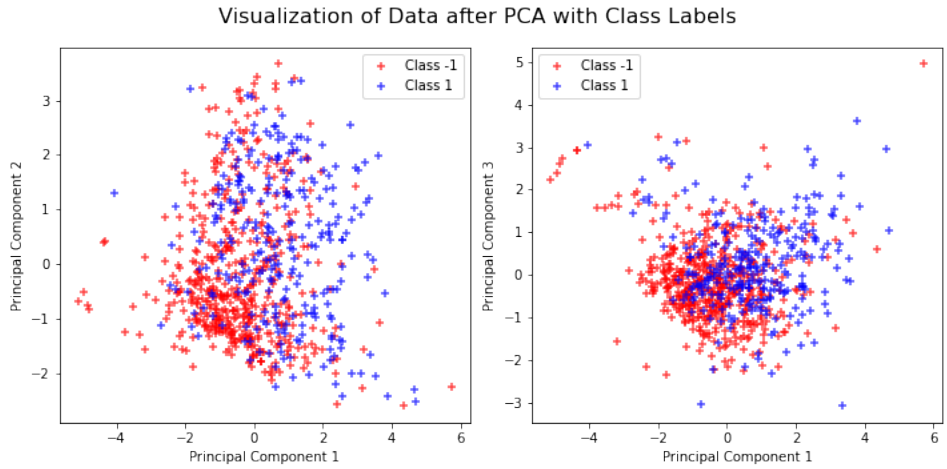

Figure 2: PCA Visualization of the Pima Indians Diabetes Database

The visualization clearly shows that the two clusters are not linearly separable in either of the displayed PCA projections. This observation suggests that employing a non-linear kernel in the SVM model could potentially improve its classification accuracy.

Training History

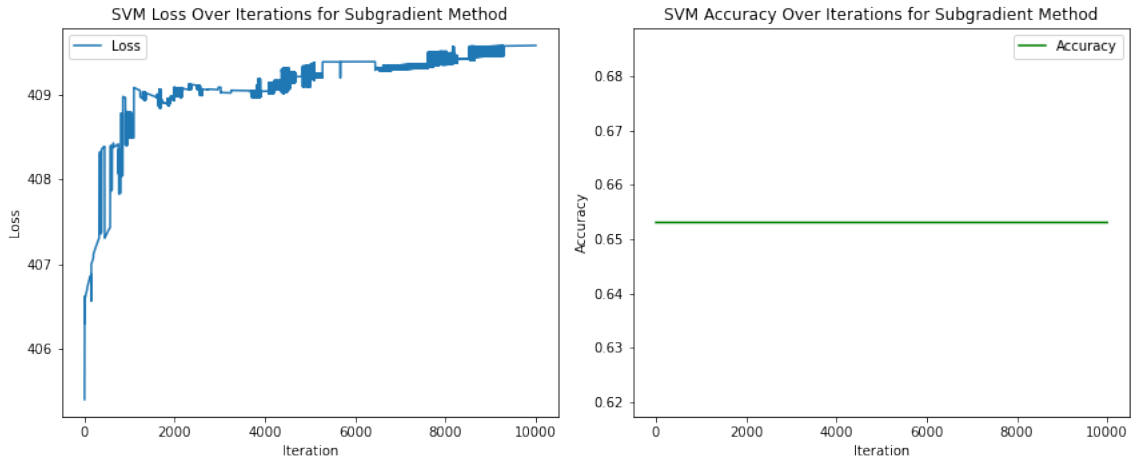

Figure 3: Training History of the Subgradient Method with 10,000 Iterations on 80% of the Pima Indians Dataset

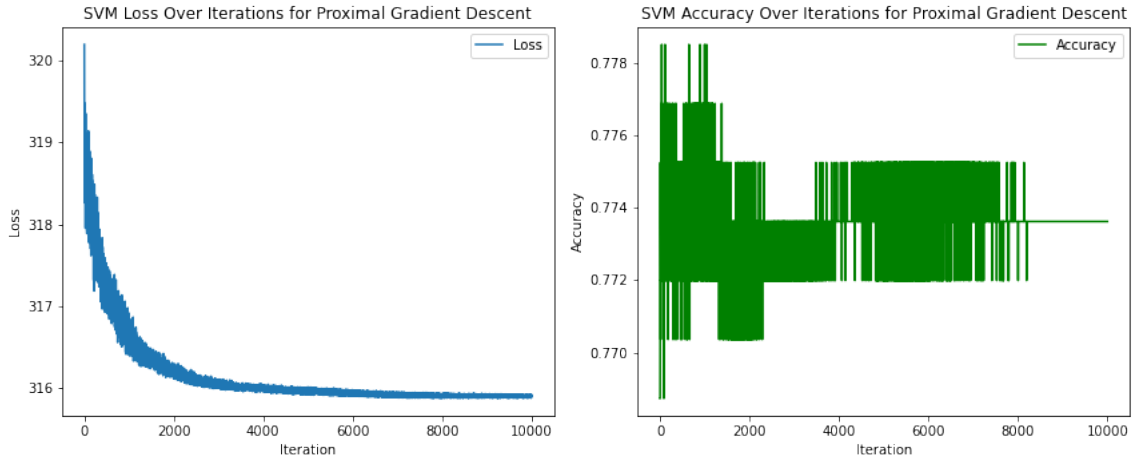

Figure 4: Training History of Proximal Gradient Descent with 10,000 Iterations on 80% of the Pima Indians Dataset

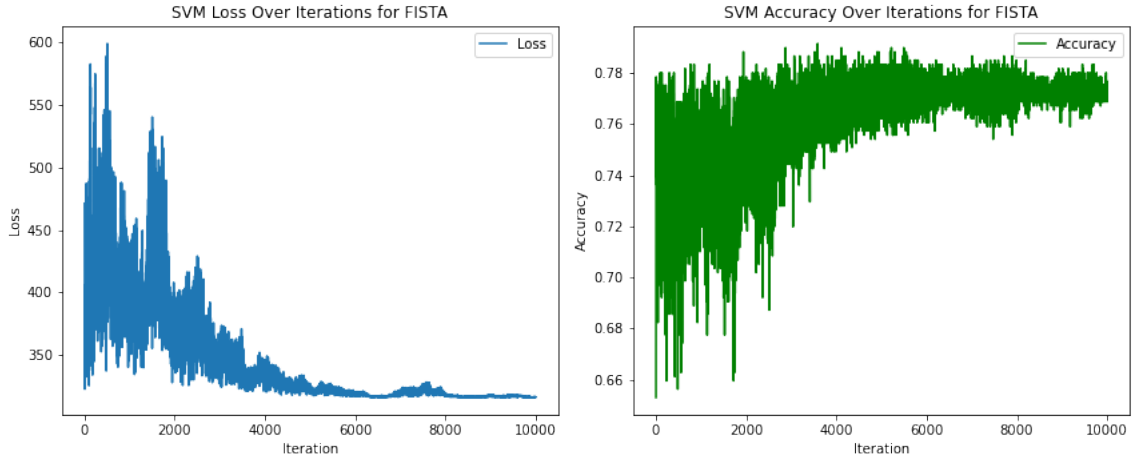

Figure 5: Training History of FISTA with 10,000 Iterations on 80% of the Pima Indians Dataset

References

[1] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2**, 1, 183–202. https://doi.org/10.1137/080716542.

[2] Chierchia, G., Pustelnik, N., Pesquet, J.-C., and Pesquet-Popescu, B. (2015). A proximal approach for sparse multiclass svm.

[3] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20**, 3 (Sep), 273–297. https://doi.org/10.1007/BF00994018.

[4] Eitrich, T. and Lang, B. (2006). Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of Computational and Applied Mathematics* **196**, 2, 425–436. https://www.sciencedirect.com/science/article/pii/S0377042705005856.

[5] Li, B., Wang, Q., and Hu, J. (2009). A fast svm training method for very large datasets. In *2009 International Joint Conference on Neural Networks*. 1784–1789.

[6] Parikh, N. and Boyd, S. P. (2014). Proximal algorithms. *Foundations and Trends in Optimization* **1**, 3, 123–231. https://web.stanford.edu/ boyd/papers/prox$_a$*lgs.html*.

[7] UCI Machine Learning and Kaggle. (2016). Pima Indians Diabetes Database. https://www.kaggle.com/uciml/pima-indians-diabetes-database. Accessed: 2023/11/29.

[8] Wang, H., Shao, Y., Zhou, S., Zhang, C., and Xiu, N. (2022). Support vector machine classifier via $l_{0/1}$l0/1 soft-margin loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 10, 7253–7265.

[9] Xu, Y., Akrotirianakis, I., and Chakraborty, A. (2015). Proximal gradient method for huberized support vector machine. *Pattern Analysis and Applications* **19**, 4 (May), 989–1005. http://dx.doi.org/10.1007/s10044-015-0485-z.