Comparison of Gradient Descent, Heavy Ball Gradient Descent Nesterov Accelerated Gradient Descent algorithms for binary classification.

A Suresh Varma
21848
varmaa@iisc.ac.in

I BV Lakshmana Raju
21190
bhargavav@iisc.ac.in

1. Introduction

1.1. Problem Statement:

The objective of this project is to compare convergence rates of various gradient based optimization methods for a binary classification problem.

1.2. Dataset:

We considered the Diabetes.csv dataset with 750 Samples each with 8 Features.

1.3. Algorithms Used:

Gradient Descent, Heavy Ball Gradient Descent, Nesterov Accelerated Gradient Descent.

1.4. Loss Function:

The loss function is L_2 regularized cross entropy loss,

$$f(w) = \frac{-1}{m}\sum_{i=1}^{m} y_i x_i^T w - \log(1+\exp(x_i^T w)) + \frac{\lambda}{2}\|w\|^2$$

Here the function is L-smooth with,

$$L = \lambda + \max_{i=1}^{m} \frac{||x_i||^2}{4}$$

and μ strongly convex with, $\mu = \lambda$.

Proof:
We have,

$$f(w) = \frac{-1}{m}\sum_{i=1}^{m} y_i x_i^T w - \log(1+\exp(x_i^T w)) + \frac{\lambda}{2}\|w\|^2$$

$$\nabla f(w) = \frac{-1}{m}\sum_{i=1}^{m} x_i(y_i - \sigma(x_i^T w)) + \lambda w$$

$$\nabla^2 f(w) = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T \sigma(x_i^T w)(1-\sigma(x_i^T w)) + \lambda I$$

For $f(w)$ to be L-smooth the maximum eigen value of the Hessian of the function must be upper bounded by L.
i.e., $\sigma(x_i^T w)(1-\sigma(x_i^T w)) \leq \frac{1}{4}$

$$\begin{aligned} L &= \lambda_{max}(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T \sigma(x_i^T w)(1-\sigma(x_i^T w)) + \lambda) \\ &\geq \lambda_{max}(\frac{1}{4}\sum_{i=1}^{m} x_i x_i^T) + \lambda \\ &= \frac{1}{4}\lambda_{max}(\sum_{i=1}^{m} x_i x_i^T) + \lambda \\ &= \frac{1}{4}\max ||x_i||_2^2 + \lambda \end{aligned}$$

$\therefore$ For the given dataset L=18.1 after normalization.

For $f(w)$ to be μ strongly convex the minimum eigen value of the Hessian of the function must be lower bounded by μ.

$$\begin{aligned} \mu &= \lambda_{min}(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T \sigma(x_i^T w)(1-\sigma(x_i^T w)) + \lambda) \\ &\leq \lambda_{min}(\frac{1}{4}\sum_{i=1}^{m} x_i x_i^T) + \lambda \\ &= \frac{1}{4}\lambda_{min}(\sum_{i=1}^{m} x_i x_i^T) + \lambda \\ &= 0 + \lambda \\ &= \lambda \end{aligned}$$

$\therefore$ For the given dataset $\mu = \lambda$.

2. Algorithms

2.1. Update Equations

i) Gradient Descent

$$w_{t+1} = wt - \eta.\nabla w_t$$
$$\eta = \frac{1}{L}$$

ii) Heavy ball Gradient Descent

$$\begin{aligned} Update_t &= \beta \times Update_{t-1} + \eta \times \nabla w_t \\ w_{t+1} &= w_t - Update_t \\ \eta &\in (0, \frac{2}{L}) \\ 0 \leq \beta &< \frac{1}{2}\left(\mu\frac{\eta}{2} + \sqrt{\frac{\mu^2\eta^2}{4} + 4(1-\eta\frac{L}{4})}\right) \quad [2] \end{aligned}$$

iii) Nesterov Accelerated Gradient Descent

$$\begin{aligned} y_{t+1} &= w_t + \eta_t \times \nabla w_t \\ z_{t+1} &= z_t - \eta_t \left(\frac{t+1}{2}\right) \nabla w_t \\ w_{t+1} &= \left(\frac{t+1}{t+3}\right) y_{t+1} + \left(\frac{2}{t+3}\right) z_{t+1} \\ \eta &= \frac{1}{L} \end{aligned}$$

iii) Intuitive Nesterov Accelerated Gradient Descent

$$\begin{aligned} w_{lookahead} &= w_t - \beta \times Update_{t-1} \\ Update_t &= \beta \times Update_{t-1} + \eta \times \nabla w_{lookahead} \\ w_{t+1} &= w_t - Update_t \\ \eta &= \frac{1}{L} \quad [1] \end{aligned}$$

2.2. Theoretical Convergence rates of L-smooth and μ strongly convex function for ϵ accuracy

i) Gradient Descent

$$T \simeq k.O(\log(\frac{1}{\epsilon})$$

ii) Heavy ball Gradient Descent i) Gradient Descent

$$T \simeq \sqrt{k}.O(\log(\frac{1}{\epsilon}) \quad [2]$$

iii) Nesterov Accelerated Gradient Descent

$$T \simeq \sqrt{k}.O(\log(\frac{1}{\epsilon})$$

3. Results and Observations

3.1. Iterations and Plots

ϵ	GD	HBGD	NAGD	Intuitive NAGD
1e-3	209	47	13	36
1e-4	326	58	15	54

Table 1. No. of iterations vs. accuracy

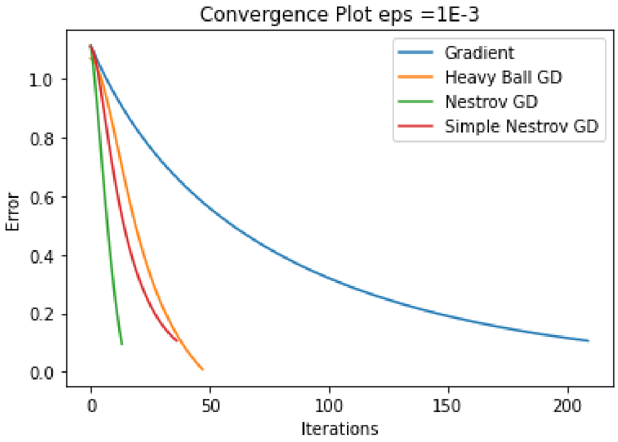

Figure 1. Error vs. Iterations for $\epsilon = 1e-3$

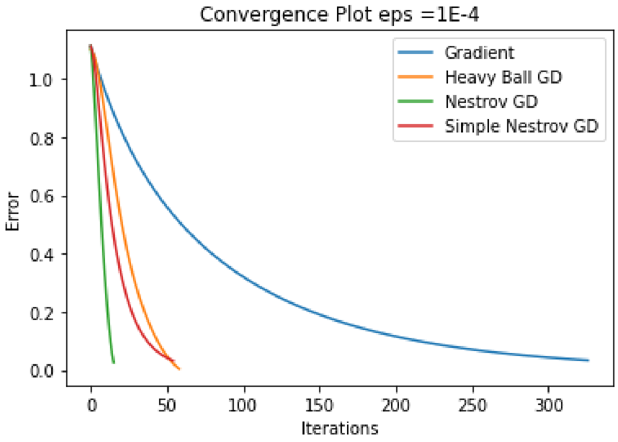

Figure 2. Error vs. Iterations for $\epsilon = 1e-4$

3.2. Observations

- Initially we applied Gradient Descent algorithm on the L_2 regularized loss function to obtain optimum loss value and the corresponding optimum weights. As the loss function is convex the obtained optimum value is global minimum.

- The stopping criteria used is, At each iteration when the absolute difference between the current loss value and optimum value is less than ϵ (1e-3, 1e-4) then the algorithm is stopped.

- As expected the Nesterov accelerated Gradient Descent converge faster compared to the Gradient descent and Heavy ball Gradient Descent which is justified because of the added momentum and quick correction in the momentum value compared to the heavy ball method.

- The practical convergence rates differ from the theoretical analysis which maybe due to ill-conditioned condition number ($k \approx 180$) and imbalance in dataset.

4. References

[1] Ilya Sutskever, PhD thesis Section 7.2.
http://www.cs.utoronto.ca/~ilya/pubs/ilya_sutskever_phd_thesis.pdf

[2] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, Mikael Johansson. Global convergence of the Heavy-ball method for convex optimization.
https://arxiv.org/abs/1412.7457