Comparative Study of Stochastic Gradient Descent Variants with Correlation Effects

Ahmed Abdalla [1,†], **Najlaa Moahmed**[1] **and Jeremie Nlandu**[1]

[1]*African Institute for Mathematical Sciences, Senegal. E-mails: aabdalla@aimsammi.org, nmohamed@aimsammi.org , jeremie@aimsammi.org,*

Abstract

This report presents an implementation of gradient methods to solve the ridge regression problem under varying data correlation conditions. The study explores the differences among stochastic gradient variants, examining how each method mitigates noise and impacts overall convergence.

Keywords: Gradient descent, Stochastic Gradient Descent, Ridge Regression

1. Introduction

Gradient descent is widely used in machine learning to minimise model loss functions. However, due to its computational intensity, a more efficient algorithm, stochastic gradient descent (SGD), is often preferred. Unlike gradient descent, SGD updates model parameters using individual samples or small batches, which significantly reduces computational load specially for large datasets[1].

SGD with a constant step size struggles with noise in gradient estimates, causing the algorithm to oscillate around the minima and preventing precise convergence .
To mitigate this issue, variance reduction techniques such as SVRG [2] and SAGA [3], along with optimized step size strategies, are employed.

This project evaluates the performance and convergence of various SGD variants and traditional gradient descent methods with different step size settings. The methods are tested on two synthetic datasets: one with high correlation and one with low correlation among features, and applied to the solve the ridge penalized empirical risk minimization for regression.

2. Problem statement

Regression and classification are both the main tasks in supervised learning. Very often, the goal of these tasks consists of solving an optimization problem. Namely, we want to solve the minimization problem

$$w^* = \arg\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell(x_i^\top w, y_i) + \frac{\lambda}{2}\|w\|_2^2 \right\} \tag{1}$$

where l is the loss function, λ a penalty constant, and the pairs (x_i, y_i) are the training points. The problem eq.1 is said :

- Rigde Regression if the loss l is quadratic, i.e $l(u,v) = (u-v)^2$,
- Rigde Logistic Regression if the loss l is a logistic loss, i.e $l(u,v) = log(1+\exp(-uv))$.

Depending on the form of the loss function l this problem can be unnameable. That is it can be difficult to obtain a closed form solution analytically. Iterative methods, like Gradient descents [1], can efficiently be used to approach the solution.

3. Stochastic Gradient Descent

The Gradient descent searches for the optimal soulution by following the steepest step by evaluating the full gradient.The fig 1 describe how this method proceed. Gradient Descent suffers from the fact that it requires to evaluate the full gradient $\nabla f(w)$ of the entire dataset at each iteration.
The SGD uses only the gradient of a single data function $\nabla f_i(w)$ at each iteration, by considering an unbiased estimate of the full gradient. The Fig 1 describes the SDG, as well.

Algorithm GD

starting points $w_0 \in \mathbb{R}^d$, **learning rate** $\alpha > 0$
for $k = 0, 1, 2, \cdots, T-1$ **do**
  $w_{k+1} = w_k - \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(w_k)$
end for
Output w_T

Algorithm SGD : Constant step size

starting points $w_0 \in \mathbb{R}^d$, **learning rate** $\alpha > 0$
for $k = 0, 1, 2, \cdots, T-1$ **do**
  Sample $j \in \{1, \ldots, n\}$
  $w_{k+1} = w_k - \alpha \nabla f_i(w_k)$
end for
Output w_T

Figure 1. Gradient Descent and Stochastic Gradient Descent algorithms

4. Experiments

4.1. Data Generation

We used a synthetic data with 50 features and 1000 samples was generated, using ground truth coefficients defined by

$$w_{\text{model\_truth}} = (-1)^{\text{idx}} \cdot \exp\left(-\frac{\text{idx}}{10}\right) \tag{2}$$

Two datasets were created: one with a high correlation of 0.7 and the other with a low correlation of 0.1. Both datasets were used in linear regression to compare behaviors.

4.2. Choice of Model and Regularization

Linear Regression (LinReg) and Logistic Regression (LogReg) models both used regularization parameter $\lambda = \frac{1}{\sqrt{n}}$ to control complexity and improve generalization.

4.3. Gradient Checking

Gradient checking was performed to validate the gradient descent implementation.
This involved comparing analytical gradients with numerical approximations, with the mean error computed and plotted to ensure correctness. Additionally, The LBFGS method was used to find accurate solutions for the models. This provided a benchmark for comparing the performance of SGD and GD.

4.4. Experiments Results

In the experimental results, we conducted comparisons between methods using synthetic data with both high correlation ($\rho = 0.7$)

and low correlation ($\rho = 0.1$).
High correlation increases noise in gradient estimates, leading to more oscillations and instability leading to the methods that are better at managing noise, such as those with shrinking step sizes or momentum, performing better and achieve smoother convergence. In contrast, low correlation reduces noise, resulting in smoother and more effective convergence across all methods.

1. **Comparing High and Low Correlation Data Performances of Constant and Shrinking SGD**: In Low correlation setup, the noise in gradient estimates is not large, improving performance for both SGD variants. However, the shrinking stepsize variant still outperforms the constant stepsize, providing smoother and more effective convergence. In contract, High correlation setup leads to higher errors for SGD with constant stepsizes, while the shrinking stepsize variant performs better and results in smoother convergence.

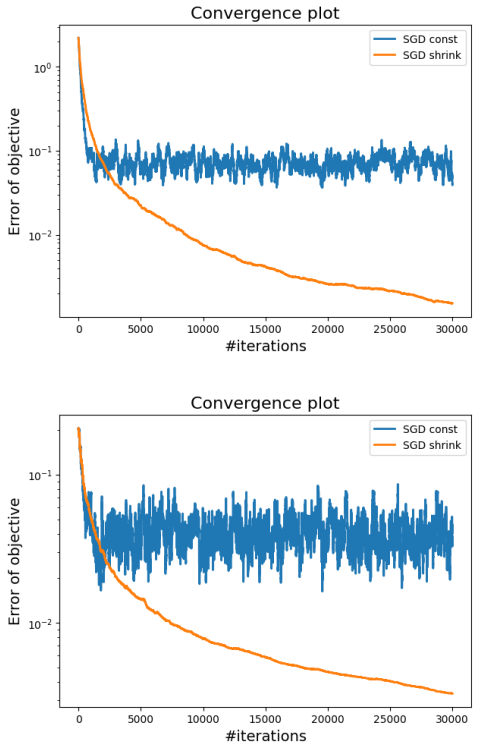

Figure 2. Constant step and shrinking steps SGD: plots of the errors of the objective vs iterations for 1) low correlated data and 2) higher correlated data.

2. **Comparing Shrinking Stepsizes and switch to shrink stepsizes:** This comparison shows (see fig. 3) that in low correlation setup, both SGD variants perform well, but the shrinking stepsize variant provides smoother convergence with fewer oscillations. The switch variant shows better overall performance but with more initial variability.
   In high correlation data, both variants are adversely affected, causing more oscillations. However, the shrinking stepsize variant manages the noise better, resulting in smoother and more stable convergence.
3. **Comparing switch to shrinking step sizes to SGD with averaging:** In low correlation setup, the averaging variant provides smoother convergence with fewer oscillations, while the switch variant performs well but starts with more variability, but ultimately achieving better performance.
   In high correlation data, both are affected by noise, but the averaging variant results in smoother and more stable convergence, as shown in fig. 4.

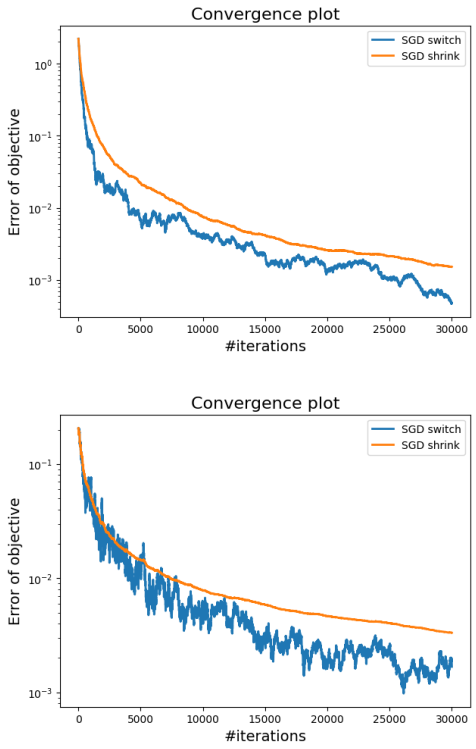

Figure 3. Shrinking and switching steps SGD: plots of the errors of the objective vs iterations for 1) low correlated data and 2) higher correlated data.

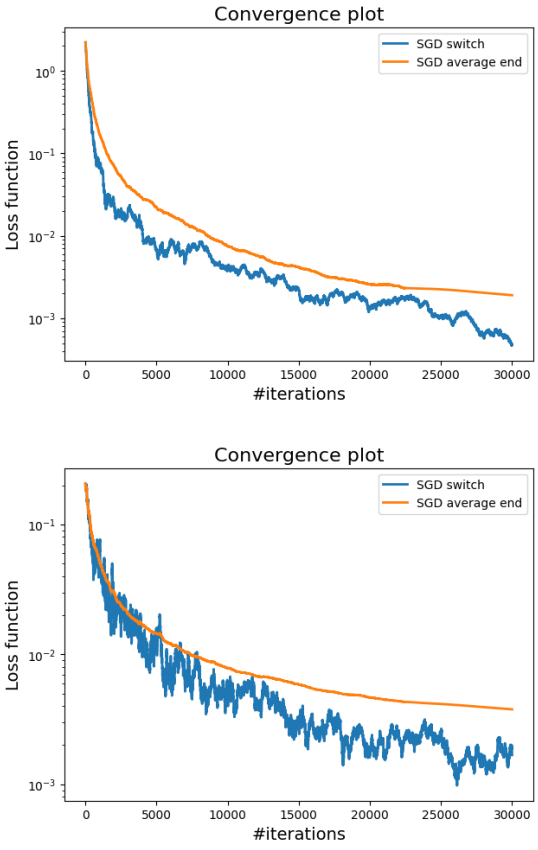

Figure 4. Switch to shrinking steps SGD with averaging: plots of the errors of the objective vs iterations for 1) low correlated data and 2) higher correlated data.

4. **Comparing all variants:** In low correlation setup, the SGD switch variant performs well initially, but SGD with momentum ($\beta = 0.4$) ultimately provides the best performance, with the lowest error and most stable convergence.
   In high correlation setup, all methods are impacted, but SGD with momentum ($\beta = 0.6$) outperforms the others by achieving the lowest loss and distance to the minimum, effectively managing the noise introduced by correlated features.

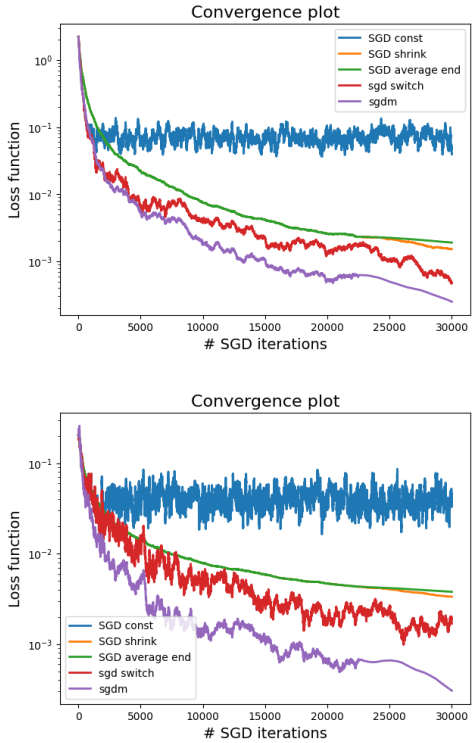

Figure 5. Comparison of all variants of the SGD: plots of the errors of the objective vs iterations for 1) low correlated data and 2) higher correlated data.

5. **Comparing full gradient descent to sgd with momentum and switch variants:** This comparison (see fig 6) shows that in low correlation setup , Gradient descent performs best overall, smoothly converging to very low loss values and effectively reaching the minima. SGD momentum with $\beta = 0.4$ performs well, surpassing SGD switch in later stages with lower loss and better convergence, but only Gradient Descent effectively reaches the minima.
   In high correlation setup, SGD momentum with $\beta = .6$ outperforms the other variants by achieving the lowest loss and distance to the minimum. Gradient Descent, while smooth, is slower due to high correlation noise, and SGD switch struggles with oscillations. Overall, all methods perform relatively similarly in the presence of high correlation.

5. Insights and Conclusion

This project investigated various stochastic gradient methods on two datasets with different correlation levels, focusing on linear regression. The experiments highlighted that high correlation introduces more noise and instability in gradient estimates, challenging convergence. In such cases, methods like shrinking stepsizes or momentum demonstrated better performance by managing noise effectively. Low correlation, on the other hand, resulted in reduced noise, facilitating smoother and more efficient convergence across all methods. The insights gained from this study emphasize the importance of selecting appropriate gradient strategies based on data correlation characteristics to optimize regression outcomes.

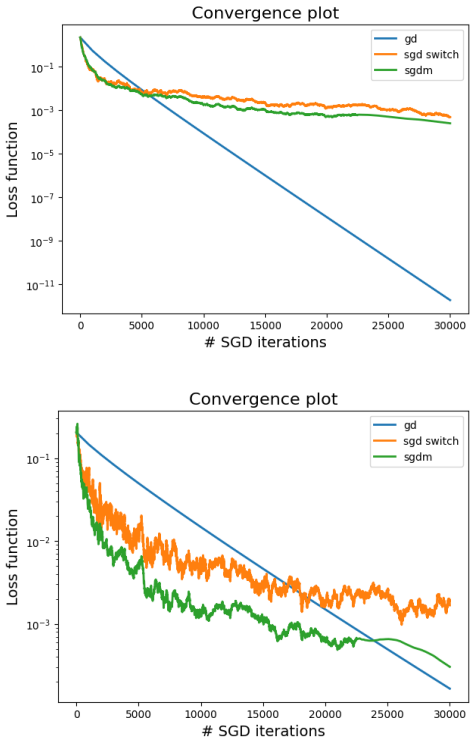

Figure 6. Comparison of the full gradient descent with the SGD with momentum and switch variants: plots of the errors of the objective vs iterations for 1) low correlated data and 2) higher correlated data.

References

[1] L. Bottou, "Large-scale machine learning with stochastic gradient descent", in *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, Springer, 2010, pp. 177–186.

[2] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction", *Advances in neural information processing systems*, vol. 26, 2013.

[3] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives", *Advances in neural information processing systems*, vol. 27, 2014.