Final Project

Article: "Statistical analysis of stochastic gradient methods for generalized linear models" by Toulis et al.
Collaborators: None

1 Introduction

The stochastic gradient descent or SGD is an iterative algorithm that finds an estimate of the optimum parameter of a loss function. Each iteration of the standard SGD procedure directly computes the estimate using the previous estimate. Despite the general applicability of the algorithm, it can be numerically unstable because it is sensitive to the choice of learning rate or step size. That is the mean squared error can erratically deviate due to learning rate misspecification. Thus, an alternative algorithm that is more robust to learning rate specification is much more appealing. One such alternative algorithm is given attention in the article, "Statistical analysis of stochastic gradient methods for generalized linear models"[5] by Toulis et al. This alternative SGD method has the current estimate on both sides of the update equation, so it has an "implicit" nature. Thus, it is known as implicit SGD or ISGD. The paper suggests that the implicit SGD can be a "competitive method for large-scale machine learning tasks"[5]. However, the authors do not run any experiments on large datasets, compare the computational cost, or compare ISGD to adaptive learning SGD methods. Thus, for this project, we investigate the performance of implicit SGD in classifying different large datasets.

For this project, we compare classification models made using SGD, ISGD, and ADAM. We compare the average computational cost, rate of convergence, asymptotic bias of the estimate, stability, and asymptotic variance of the estimate. We find that all the methods have the same asymptotic bias. The SGD method has the lowest computational cost, whereas ISGD has the highest and is double that of SGD. The ADAM method has the fastest rate of convergence and tends to be the most stable, but also tends to have the highest asymptotic variance. Overall, ADAM is the most competitive method out of the three methods.

2 Background

Let $y \in \mathbb{R}$ be the observation, $x \in \mathbb{R}^p$ be the vector of features and $\theta^* \in \mathbb{R}^p$ be the unknown model parameters. Then, $\ell(\theta; y_n)$ is the log-likelihood of θ given the observation y_n at some iteration n. The SGD algorithm updates an estimation θ_n. It does so by first randomly and uniformly selecting a small batch of observations for each iteration. Then, it updates each iteration n by a learning rate a_n:

$$\theta_n \leftarrow \theta_{n-1} - a_n \nabla \ell(\theta_{n-1}; y_n).$$

On the other hand, the ISGD algorithm has the update equation:

$$\theta_n \leftarrow \theta_{n-1} - a_n \nabla \ell(\theta_n; y_n).$$

In order to compute this update, we start with an initial guess for the left value of θ_n, such as θ_{n-1}. Then, we repeatedly update the parameter using the equation above until the equality is satisfied. In practice, we can use a root-finding method to solve the update equation. Due to the implicit nature of the update, the computational cost for this method can be high.

The article by Toulis et al[5] contains exact equations for the asymptotic bias, stability, and asymptotic variance. Let the bias of the estimate be $b_n = \mathbb{E}(\theta_n) - \theta^*$, the identity matrix be $\boldsymbol{I}$, the dispersion parameter which affects the variance of the outcome be ψ, and $\mathcal{C}(a_n)$ be an arbitrary a_n-convergent series. Also, for some $\alpha > 0$, the ratio $a_{n-1}/a_n = 1 + (1/\alpha)a_n + \mathcal{O}(a_n^2)$. If this assumption holds then the asymptotic bias of SGD is

$$b_n = [\boldsymbol{I} + a_n \psi \mathbb{E}(\nabla \nabla \ell(\theta^*; y, x))] b_{n-1} + \mathcal{C}(a_n).$$

Whereas the asymptotic bias of ISGD is

$$b_n = [\boldsymbol{I} - a_n \psi \mathbb{E}(\nabla \nabla \ell(\theta^*; y, x))]^{-1} [b_{n-1} + \mathcal{C}(a_n)].$$

From these equations we can see that the SGD converges faster than ISGD because $\|\boldsymbol{I} + a_n \psi \mathbb{E}(\nabla \nabla \ell(\theta^*; y, x))\|_2 < \|[\boldsymbol{I} - a_n \psi \mathbb{E}(\nabla \nabla \ell(\theta^*; y, x))]^{-1}\|_2$. Also, as n approaches the limit, both methods have the same convergence rate and are unbiased. If we ignore the remainder term of the bias equations, then we can denote the bias of SGD to be $P_1^n b_0$ and ISGD to be $Q_1^n b_0$, where b_0

is the initial bias of the common initial parameter θ_0. Now, we can find the maximum eigenvalue of SGD which is

$$\max\{\text{eig}(P_1^n)\} = \Theta\left(2^{\alpha\psi\lambda_{(p)}}/\sqrt{\alpha\psi\lambda_{(p)}}\right)$$

and for ISGD which is

$$\max\{\text{eig}(Q_1^n)\} = \mathcal{O}(1).$$

These equations tell us the stability of these methods. For SGD the effect from the initial conditions, such as the learning rate, can be exponentially increased before receding. Whereas, for the implicit case the initial conditions monotonically decrease. Thus, ISGD is more robust to learning rate misspecification.
Finally, the asymptotic variance of the estimate for both SGD and ISGD is,

$$(1/a_n)\,\text{Var}(\theta_n) \to -\alpha\psi^2(-2\alpha\psi\mathbb{E}(\nabla\nabla\ell(\theta^*;y,x)) - \boldsymbol{I})^{-1}\mathbb{E}(\nabla\nabla\ell(\theta^*;y,x)).$$

So, both methods have the same asymptotic variance.

From the equations by Toulis et al we find that SGD has a faster convergence rate and ISGD is more stable but at the limit both methods have the same convergence rate, bias, and variance. The authors carried out a set of experiments to validate these findings. However, the experimental results are lacking. They do not compare the computational cost between the methods, even though it is apparent that the implicit nature of ISGD makes it quite inefficient. As mentioned before, the authors claim that ISGD is a "competitive method for large-scale machine learning tasks"[5] even though they do not conduct any tests on large datasets. Furthermore, they do not compare the methods with adaptive learning methods to find out how competitive ISGD is. Therefore, this paper addresses these three limitations by investigating the performance of ISGD in classifying different large datasets. For the third limitation, this paper compares the SGD and ISGD methods with the state-of-the-art adaptive learning method, ADAM [3]. The ADAM optimizer builds upon the SGD algorithm by using exponential moving averages to adjust the learning rate for each parameter to achieve stability.

3 Experiments

For the empirical investigation, a classification model is built for the three optimizers: SGD, ISGD, and ADAM. The datasets that are used are the Titanic [2], MNIST [4], and Iris [1] datasets. The Titanic dataset contains data for the passengers that were on the RMS Titanic, and for this dataset, a binary logistic regression model is used to classify the passengers that survived the shipwreck. The MNIST dataset consists of handwritten digits and a binary logistic regression model is used to classify the number 8. Finally, the Iris dataset contains data for three different types of iris' and a multinomial logistic regression model is used to classify each flower. For all the datasets, 80% of the data is used for the training sets and the rest for the testing sets. To compute the bias and variance of the models, the average values are taken over five training sets. Unless stated otherwise, the fixed experimental conditions are $N = 1000$ iterations, and $a = 0.01$ constant learning rate. For the ADAM optimizer, the default settings as stated in the original paper are used [3].

3.1 Average computational cost

As stated previously, the implicit nature of ISGD can result in a high computational cost. Thus, we compare the average computational cost of building a model for the three methods. To find the average computational cost for a model, the time to build five models is recorded using the time.process_time() function.

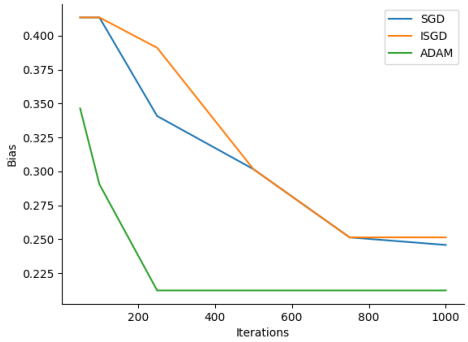
Method	Titanic	MNIST	Iris
SGD	0.0	26.0	0.05
ISGD	0.122	38.9	0.106
ADAM	0.003	32.7	0.072

Table 1: Average computational cost in seconds.

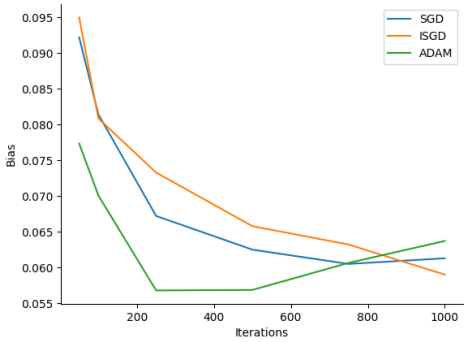
From Table 1, we can see that for all three datasets, the SGD models have the lowest computational cost, and ISGD has the highest. This supports the theory since SGD has the simplest update equation. The cost of ISGD is more than double of SGD for some datasets, so it is not quite attractive, especially for large-scale tasks that have long runtimes.

3.2 Rate of convergence and asymptotic bias

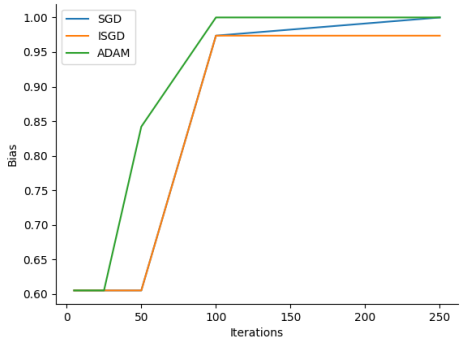
In the next few experiments, we check if the theoretical results hold for the datasets and compare them with the ADAM optimizer. In this experiment, we find both the rate of convergence and asymptotic bias for the models by investigating the effect that the number of iterations has on the bias. The set of iterations that we run the experiment for is $N = \{50, 100, 250, 500, 750, 1000\}$, except for the Iris model which is $N = \{5, 10, 25, 50, 100, 250\}$ because it converges much faster.



(a) Titanic Dataset



(b) MNIST dataset



(c) Iris Dataset

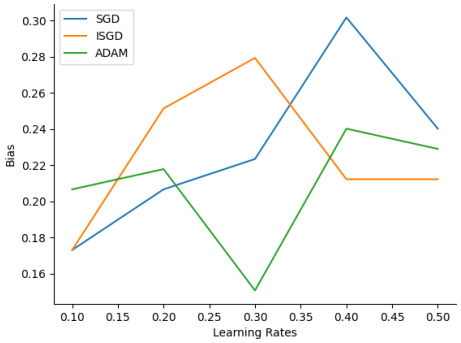
Figure 1: Bias over iterations.

In Figure 1, for the Titanic and MNIST datasets the bias decreases as the number of iterations increases, but for the Iris dataset, the bias increases

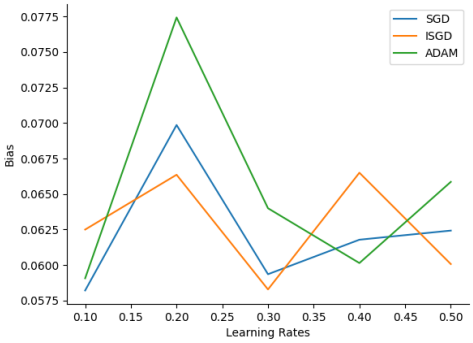
instead. This could be because the Iris dataset is the only one that is classified with the multinomial logistic regression model. The models are only asymptotically unbiased for the Titanic and MNSIST datasets. The value of the bias at 1000 iterations is similar for all the SGD and ISGD models, except for the model of the Titanic dataset using the ADAM method which is lower in comparison. We can see that the ADAM models have the fastest rate of convergence, and the ISGD has the slowest rate. The SGD and ISGD models have similar rates of convergence especially at higher iterations.

3.3 Stability

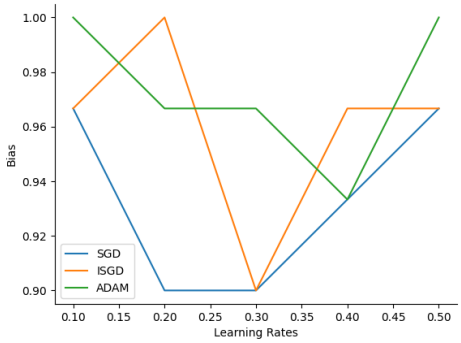
We investigate the stability of the models by checking the effect that the learning rate has on the bias. The set of learning rates that is used is $a = \{0.1, 0.2, 0.3, 0.4, 0.5\}$.



(a) Titanic Dataset



(b) MNIST dataset



(c) Iris Dataset

Figure 2: Bias over learning rate.

The variance of the bias values gives us the stability, that is the higher the variance of the bias, the more unstable it is. In Figure 2, for the Titanic dataset, the ADAM method is the most stable and SGD is the least. As for the MNSIT dataset, ISGD is the most stable and ADAM is the least. Finally, for Iris, ADAM is the most stable and ISGD is the least. It is hard to determine which algorithm is most stable due to inconsistency, but the ADAM optimizer does tend to be the most stable. Perhaps to resolve this issue, the bias values should be taken over a larger number of training sets.

3.4 Asymptotic Variance

In this experiment, we find the asymptotic variance for the models.

Method	Titanic	MNIST	Iris
SGD	0.023	0.014	0.016
ISGD	0.024	0.016	0.016
ADAM	0.012	0.051	0.074

Table 2: Variance for 1000 iterations.

From Table 2, we can see that the asymptotic variance for the SGD and ISGD models are the same or very close. So, the theoretical result was held. For ADAM, the asymptotic variance is larger for the MNIST and Iris datasets but is smaller for the Titanic dataset.

4 Conclusion

All the models are asymptotically unbiased, SGD has the lowest computational cost, ADAM has the fastest rate of convergence and tends to be the most stable, and both ISGD and SGD tend to have the lowest asymptotic variance. Overall, the most competitive method is the ADAM method. For future experiments, it would be better to find the bias over a larger number of training sets because the stability results are a bit inconclusive. Also, it would be good to test the results on more different types of models such as SVM or neural networks.

References

[1] R.A. Fisher. iris, 2018.

[2] Thomas Cason Frank E. Harrell Jr. Titanic dataset, oct 2017.

[3] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[4] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[5] Panagiotis Toulis, Edoardo Airoldi, and Jason Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 667–675, Bejing, China, 22–24 Jun 2014. PMLR.