EMPIRICAL STUDY OF VARIANCE-REDUCED METHODS FOR MACHINE LEARNING

Sathvik Joel K
Dept. of Computer Science and Engineering
IIT Madras
cs19b025@smail.iitm.ac.in

Vishnu Vinod
Dept. of Computer Science and Engineering
IIT Madras
cs19b048@smail.iitm.ac.in

***K*eywords** Gradient Descent · Variance Reduction · SVRG · SAG · SDCA · ADAM

1 Introduction

The literature on variance reduced (VR) stochastic gradient algorithms has extensively studied their iterative bounds[1]. However, there has been a noticeable lack of empirical studies evaluating and comparing the practical performance of these algorithms on both convex and non-convex problems, and across various datasets. Moreover, many of the proposed algorithms lack sufficient empirical evaluations in their original papers.

In this work, we aim to address this gap by conducting a thorough empirical analysis of several VR stochastic gradient algorithms. We explore their performance in diverse problem settings, including both convex and non-convex cases, and test them on multiple datasets. We hope that our study provides valuable insights into the practical effectiveness of these algorithms and helps to bridge the gap between theory and practice. Additionally, we aim to fill the gap in the literature by providing empirical evaluations for many algorithms that lack sufficient experimental results in their original papers.

2 Related Work

Stochastic gradient descent (SGD) is a widely used optimization algorithm for training deep neural networks, but its convergence can be slow due to high variance in stochastic gradients. To address this issue, several variance reduction methods have been proposed by various authors, in the past decade, including Stochastic Average Gradient (SAG)[2][3], Stochastic Variance Reduced Gradient (SVRG)[4], and Stochastic Dual Coordinate Ascent (SDCA)[5], among others.

Initial experimental results show that all these methods can achieve faster convergence and better generalization performance compared to traditional SGD. SAG is memory-based and can reduce variance in gradients by keeping a history of past gradients. SVRG reduces bias by periodically computing the full gradient. SDCA updates one dual variable at a time and uses a cyclic coordinate descent scheme to minimize the primal objective function.

Variance reduced methods suffer from significant limitations too. SVRG and related methods have shown promising results in various tasks, such as image classification, but require careful implementation and tuning of hyperparameters. It is more versatile than most other VR methods and a number of variations/improvements of SVRG have spawned over the past decade. In a recent study by Defazio et al.[6], it was shown that the Stochastic Variance Reduced Gradient (SVRG) algorithm exhibits an increase in variance for a majority of each epoch in deep neural networks such as ResNet and DesNet. This finding highlights a potential limitation of SVRG in the context of deep learning.

In recent years, momentum-based methods have dominated both research interest and effort in enhancing the convergence of Stochastic Gradient Descent (SGD). This acceleration in convergence has been achieved by introducing a momentum term that smooths out the gradient updates. Nesterov Accelerated Gradient (NAG)[7] is one the first such widely used momentum-based method that has shown faster convergence than traditional SGD and has been applied successfully in various deep learning tasks. Another popular momentum-based method is Adaptive Moment Estimation (Adam)[8], which can handle non-convex functions and is robust to noisy gradients.

Despite their success, momentum-based methods have limitations, such as slower convergence and poorer generalization performance in certain scenarios, including flat regions of the loss surface. Additionally, the selection of hyperparameters can significantly impact their performance, requiring careful tuning.

3 Problem Statement

This project paper aims to implement and extensively compare various variance reduced stochastic gradient algorithms, including Stochastic Average Gradient (SAG), Stochastic Average Gradient "Amélioré"(SAGA) [9], Stochastic Dual Coordinate Ascent (SDCA), and Stochastic Variance Reduced Gradient (SVRG). Additionally, we want to compare these algorithms with momentum-based approaches such as Nesterov Accelerated Gradient (NAG), Adaptive Moment Estimation (Adam), Adaptive Delta (AdaDelta)[10], and Root Mean Square Propagation (RMSProp).

To evaluate the performance of these algorithms, we will conduct experiments in various problem settings, including both convex and non-convex cases. For the convex setting, we will use Logistic Regression and L2-regularized Logistic Regression. In the non-convex setting, we plan to use Empirical Risk Minimization (ERM) problem settings with different non-convex loss functions, such as sigmoid loss and hinge loss. Additionally, we will test these algorithms on feedforward neural networks, as was done in a previous studies.

We also aim to investigate the claim made in [6] by testing the performance of SVRG on shallow neural networks. We will use a range of shallow neural network architectures and evaluate the performance of SVRG in terms of variance and convergence rate. Our experiments will enable us to gain a better understanding of the performance characteristics of SVRG on both shallow networks. Moreover, our findings will provide insights into the suitability of SVRG for practical applications, particularly in the context of deep learning[11][12].

We will use standard datasets that are readily available on the LibSVM website [13], such as adult (a9a) , web (w8a), rcv1 (rcv1.binary), MNIST, and CIFAR-10. By evaluating the performance of these algorithms across a range of problem settings and datasets, we aim to provide a comprehensive analysis of their effectiveness in practice.

References

[1] Robert Mansel Gower, Mark W. Schmidt, Francis R. Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108:1968–1983, 2020.

[2] Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[3] Mark Schmidt, Nicolas Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162, 09 2013.

[4] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[5] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(16):567–599, 2013.

[6] Aaron Defazio and Leon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[7] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[9] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives, 2014.

[10] Matthew D. Zeiler. Adadelta: An adaptive learning rate method. *ArXiv*, abs/1212.5701, 2012.

[11] Zeyuan Allen Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, 2016.

[12] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 314–323, New York, New York, USA, 20–22 Jun 2016. PMLR.

[13] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), May 2011.