Stochastic Newton with Arbitrary Sampling

Denis Shveykin **Rustem Islamov**

Abstract

The problem of minimizing the average of a large number of sufficiently smooth and strongly convex functions is ubiquitous in machine learning. Stochastic first-order methods for this problem of Stochastic Gradient Descent type are well studied. In turn, second-order methods, such as Newton, have certain advances since they can adapt to the curvature of the problem. They are also known for their fast convergence rates. But stochastic variants of Newton-type methods are not studied as good as SGD-type ones and have limitations on the batch size. Dmitry Kovalev et al proposed a method which requires no limitations on batch sizes. Our goal is to explore this method with different sampling strategies that lead to practical improvements.

Keywords Stochastic Newton, sampling strategy

1 Introduction

The problem is to minimize the empirical risk which has finite-sum structure [12]:

$$\min_{x \in \mathbb{R}^d} \left[f(x) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right] \tag{1}$$

where each f_i is assumed to have Lipschitz Hessian.

The stochastic approach is used, because n is very large and it is computationally difficult to evaluate the gradient of each f_i on each step. The Stochastic Gradient Descent (SGD) method [20] calculates the gradients of some randomly chosen f_i, which has cheap iterations compared to the deterministic algorithm. Such methods have two certain advantages. Firstly, they do not rely on the number of used f_i on a single step, which is called the batch size. Therefore, these methods can be applied even with small batches.

Secondly, there are ways to solve some common problem that arises due to the stochastic nature of the methods. The fact is that stochastic methods provide convergence to a neighbourhood of the solution only [18, 7]. Its size is proportional to the variance of the stochastic gradient. So the second advantage of SGD is presence of so-called variance-reduced methods [21, 9, 4, 8], which mitigate the mentioned effect. However, these methods' computational complexity depends on the curvature of the problem, which is called the condition number and is defined as the ratio of Lipschitzness and strong convexity parameters.

This leads to usage of second-order methods, such as the Newton method [16, 10, 6]. Taking into account the derivatives of the second order makes it possible to adjust the algorithm's step sizes to the curvature of the problem [16]. Unfortunately, much less literature has been written on this topic than about first-order methods. Some proposed algorithms need extra assumptions [10, 6] or regularization [17, 3, 2] to converge. Some works [14, 15, 5] provide stochastic quasi-Newton methods, whose computational complexity exceeds such in the variance-reduced SGD variants [9].

In addition, there are many algorithms presented [11, 1, 26, 23, 22, 25], that need large batch sizes. Particularly, the required batch size is commonly quadratically proportional to the inverse of the desired accuracy. That means that one need to evaluate a large number of f_i Hessians. And this decreases the profit gained by adding randomness into the algorithm, because these batch sizes can become as big as n.

Dmitry Kovalev, Konstantin Mishchenko and Peter Richtarik proposed [12] a simple Stochastic Newton algorithm, which can work with small batches, even with batches of size one. Their algorithm does not provide unbiased estimates but nevertheless shows good convergence. This is achieved by developing new Lyapunov functions that are specific to second-order methods. Our goal is to apply different sampling strategies to this method and explore their performance.

The basic set of sampling strategies can be taken from [19], where these strategies are applied to the Parallel Coordinate Descent methods. It is also known that so-called Importance Sampling can improve the SGD performance [24, 13], since it can reduce the stochastic gradient variance.

References

[1] Raghu Bollapragada, Richard H Byrd, and Jorge Nocedal. Exact and inexact subsampled Newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 04 2018.

[2] Nikita Doikov and Yurii Nesterov. Minimizing uniformly convex functions by cubic regularization of newton method. *Journal of Optimization Theory and Applications*, 189(1):317–339, mar 2021.

[3] Nikita Doikov, Peter Richtarik, and University of Edinburgh. Randomized block cubic Newton method. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1290–1298. PMLR, 10–15 Jul 2018.

[4] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent, 2019.

[5] Robert Gower, Donald Goldfarb, and Peter Richtarik. Stochastic block bfgs: Squeezing more curvature out of data. 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 1869–1878, New York, New York, USA, 20–22 Jun 2016. PMLR.

[6] Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtarik. Rsn: Randomized subspace newton. 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] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. Sgd: General analysis and improved rates. 2019.

[8] Filip Hanzely and Peter Richtárik. One method to rule them all: Variance reduction for data, parameters and many new methods, 2019.

[9] 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.

[10] Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Global linear convergence of newton's method without strong-convexity or lipschitz gradients, 2018.

[11] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*,

volume 70 of *Proceedings of Machine Learning Research*, pages 1895–1904. PMLR, 06–11 Aug 2017.

[12] Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtárik. Stochastic newton and cubic newton methods with simple local linear-quadratic rates. *arXiv preprint arXiv:1912.01597*, 2019.

[13] Huikang Liu, Xiaolu Wang, Jiajin Li, and Anthony Man-Cho So. Low-cost lipschitz-independent adaptive importance sampling of stochastic gradients. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2150–2157, 2021.

[14] Luo Luo, Zihao Chen, Zhihua Zhang, and Wu-Jun Li. A proximal stochastic quasi-newton algorithm, 2016.

[15] Philipp Moritz, Robert Nishihara, and Michael Jordan. A linearly-convergent stochastic l-bfgs algorithm. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 249–258, Cadiz, Spain, 09–11 May 2016. PMLR.

[16] Yurii Nesterov. *Introductory lectures on convex optimization: a basic course/ by Yurii Nesterov*. Applied optimization, 87. Kluwer Academic Publishers, Boston, 2004.

[17] Yurii E. Nesterov and Boris T. Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108(1):177–205, 2006.

[18] Lam Nguyen, PHUONG HA NGUYEN, Marten van Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. SGD and hogwild! Convergence without the bounded gradients assumption. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3750–3758. PMLR, 10–15 Jul 2018.

[19] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156:433–484, 2016.

[20] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[21] 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.

[22] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

[23] Junyu Zhang, Lin Xiao, and Shuzhong Zhang. Adaptive stochastic variance reduction for subsampled newton method with cubic regularization. *INFORMS Journal on Optimization*, 4(1):45–64, 2022.

[24] Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling, 2014.

[25] Dongruo Zhou and Quanquan Gu. Stochastic recursive variance-reduced cubic regularization methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3980–3990. PMLR, 2020.

[26] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularized newton methods. In *International Conference on Machine Learning*, pages 5990–5999. PMLR, 2018.