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. [9] 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.

1

Keywords Stochastic Newton, sampling strategy

1 Introduction

The problem is to minimize the empirical risk which has finite-sum structure [9]:

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

ввести

where each f_i is assumed to have Lipschitz Hessian.

Typically, n is very large for modern real-world problems. Thus, the stochastic approach is used because it is computationally difficult to evaluate the gradient of all f_i at each step. The Stochastic Gradient Descent (SGD) method [14] calculates the gradients of some randomly chosen f_i which leads to cheaper computation per-iteration cost compared to the vanilla Gradient Descent (GD). The analysis of SGD and its modifications is rich and well explored. Firstly, the theory of SGD-type

1. Лучше в аннотации ссылки не давать. Можно написать "Previously it was proposed…"

methods does not restrict the batch size. Therefore, these algorithms can be applied even with small batches. It is known that simple SGD converges only to a neighbourhood of the solution only [12, 4] whose size is proportional to the variance of the stochastic gradient. However, there are techniques (aka variance 2 reduction) to solve this issue. These techniques [15, 6, 2, 5] modify the update rule 3 of vanilla SGD which allows to mitigate the aforementioned effect without changing per-iteration cost. However, the main disadvantage of all gradient-type methods is that a computation 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 is a place where second-order methods such as the Newton method [11, 7, 3] come to play. Taking into account second-order derivatives it possible to adjust the algorithm's step sizes to the curvature of the problem [11]. Unfortunately, much less work has been done in the direction of stochastic Newton-type methods. Many algorithms [8, 1, 20, 17, 16, 19] require large batch sizes. In particular, the required 3 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 which sometimes can be much larger than n. [9] proposes a simple Stochastic 4 Newton algorithm, which can work with batches of any size. Their algorithms achieve local linear and super-linear convergence. 5

In practice, various sampling strategies are used for SGD-type algorithms to improve 6 further the performance. One of the most famous sampling mechanisms is so-called Importance Sampling [18, 10]. The idea is to compute gradients of the functions that have more impact on the problem. [13] studies many other sampling mechanisms. 7 We aim to analyse such strategies, but for Algorithms 1 and 2 of [9] to improve the theoretical and practical applications of algorithms. We investigate various sampling strategies supporting them with rigorously constructed experiments.

2 Problem statement

We discuss the problem of empirical risk minimization (ERM). The objective function f (1) is the average of a big number of functions f_i. In fact each f_i is 8 the loss on the i-th train data point. For example, this notation can be attributed to linear regression. In this case we try to fit the model's parameter vector x to minimize the mean squared error (MSE) on the train data.

2. also known as
3. Лучше определиться у вас рассматривается, метод или алгоритм. Алгоритм - это более формализованное понятие. Эти два термина в тексте лучше не путать
4. Лучше писать "The paper [9] proposes". В идеале, если убрать ссылки на цитируемые работы, должен остаться грамматически корректный текст
5. Чей "their"? См. пункт 3.
6. См. пункт 3
7. См. пункт 4
8. Возможно "large" здесь будет лучше

2.1 Assumptions

In our work we will assume the functions f_i to satisfy the same regularity conditions that were originally introduced in [9].

Assumption 2.1 (Strong convexity). *A differentiable function* $\phi: \ \mathbb{R}^d \to \mathbb{R}$ *is* μ*-strongly convex, where* $\mu > 0$*, if* $\forall \underbrace{, x, y \in \mathbb{R}^d}$

9

$$\phi(x) \geqslant \phi(y) + \langle \nabla \phi, x - y \rangle + \frac{\mu}{2} \|x - y\|^2, \tag{2}$$

where the norm $\|\cdot\|$ is Euclidean. For twice differentiable functions this assumption is equivalent to the Hessian having each eigenvalue $\geqslant \mu$.

Assumption 2.2 (Lipschitz Hessian). *A function* $\phi : \mathbb{R}^d \to R$ *has H-Lipschitz Hessian if* $\forall \ x, y \in \mathbb{R}^d$

$$\|\nabla^2 \phi(x) - \nabla^2 \phi(y)\| \leqslant \|x - y\| \tag{3}$$

2.2 Sampling

Definition 2.1 (Sampling). A random set-valued mapping $\hat{S} : \ [n] \to 2^{[n]}$ is called sampling.

That means that each $S_k \subseteq [n]$ is a realization of $\hat{S}$. In this case we can call any particular probability distribution on $2^{[n]}$ a sampling strategy.

2.3 Algorithm

We will be applying different sampling strategies for the Algorithm 1 from [9]:

Algorithm 1 Stochastic Newton (SN)

```
Initialize: Choose starting iterates $w_1^0, w_2^0, ...w_n^0 \in \mathbb{R}^d$

for $k = 0, 1, 2, ...$ do
    $x^{k+1} = \left[\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(w_i^k)\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(w_i^k) w_i^k - \nabla f_i(w_i^k)\right]$
    Choose a subset $S^k \subseteq \{1, 2, ..., n\}$ with one of the sampling strategies
    $w_i^{k+1} = \begin{cases} x^{k+1} & i \in S^k \\ w_i^k & i \notin S^k \end{cases}$
end for
```

9. Есть достаточно распространенная нотация -
векторы и вектор-функции обозначаются жирными буквами, матрицы - жирными заглавными. Проверить по работам по оптимизации, и в частности по работам Рустема. Если там другая нотация - ок, можно не исправлять.
Если нотация соответствует описанной - привести в порядок обозначения векторов и матриц.

2.4 Goal of the project

We aim to explore different sampling strategies for the Algorithm 1 in order to obtain practical improvements and their theoretical explainations.

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

[3] 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.

[4] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. Sgd: General analysis and improved rates. 2019.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.