On the Practical Robustness of the Nesterov's Accelerated Quasi-Newton Method

S. Indrapriyadarsini [1], Hiroshi Ninomiya [2], Takeshi Kamio [3], Hideki Asai [4]

[1] Graduate School of Science and Technology, Shizuoka University, Japan
[2] Graduate School of Electrical and Information Engineering, Shonan Institute of Technology, Japan
[3] Graduate School of Information Sciences, Hiroshima City University, Japan
[4] Research Institute of Electronics, Shizuoka University, Japan
s.indrapriyadarsini.17@shizuoka.ac.jp, ninomiya@info.shonan-it.ac.jp, kamio@hiroshima-cu.ac.jp,
asai.hideki@shizuoka.ac.jp

Abstract

This study focuses on the Nesterov's accelerated quasi-Newton (NAQ) method in the context of deep neural networks (DNN) and its applications. The thesis objective is to confirm the robustness and efficiency of Nesterov's acceleration to quasi-Netwon (QN) methods by developing practical algorithms for different fields of optimization problems.

Introduction

Optimization forms the core in several areas of engineering, statistics, machine learning, neural networks, quantum computing, fundamental sciences, etc. Hence there is a dire need for solving large scale non-linear optimization problems. A good optimization algorithm is expected to perform well across different types of problems (robustness) with reasonable computation and storage costs (efficiency) and less sensitivity to error and noise (accuracy). Gradient based algorithms have been widely used in optimization and can be categorized as (1) first order methods (eg. SGD, Adam) (2) higher order methods (eg. Newton method, quasi-Newton method) and (3) heuristic derivative-free methods (eg. coordinate descent, SPSA), each with its own pros and cons. Much progress has been made in the last 20 years in designing and implementing robust and efficient methods and yet there are many classes of applications where current state of the art optimizers fails. In the era of immense data, the effectiveness and efficiency of the optimization algorithms dramatically influence the popularization and application. To this end, this thesis aims at developing NAQ based optimizers and demonstrate its practical robustness, accuracy and efficiency for different applications in the ML/NN context.

Background

The solution to a mathematically modelled problem can be obtained by optimizing the objective function subject to contraints on its variables. Given an objective function $f(\boldsymbol{\theta}_k)$, optimization algorithms iteratively minimize (or maximize) until they terminate at the optimum solution $\boldsymbol{\theta}^*$.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f(\boldsymbol{\theta}) \tag{1}$$

First order gradient based methods have low computation complexity but exhibit slow convergence. Contrarily, second order methods such as the Newton's method converge quadratically, but incur high computation costs especially with increase in dimensionality of the problem. Thus, quasi-Newton (QN) methods (Nocedal and Wright 2006) such as SR1 and BFGS have been widely used and the iterative parameter update minimizing (1) takes the form

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \mathbf{H}_k \nabla f(\boldsymbol{\theta}_k), \tag{2}$$

where $\mathbf{H}_k$ is a symmetric positive definite matrix approximated by the following BFGS formula.

$$\mathbf{H}_{k+1}^{\mathrm{BFGS}} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^{\mathrm{T}}) \mathbf{H}_k^{\mathrm{BFGS}} (\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^{\mathrm{T}}) + \rho_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}}, \tag{3}$$

$$\rho_k = \frac{1}{\mathbf{y}_k^{\mathrm{T}} \mathbf{s}_k}, \mathbf{s}_k = \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k, \ \mathbf{y}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k). \tag{4}$$

Several studies have demonstrated the robustness of the BFGS and SR1 methods and its variants (Mokhtari and Ribeiro 2015; Berahas et al. 2019). Recently, NAQ (Ninomiya 2017) was proposed by applying the Nesterov's acceleration (Nesterov 1983) to BFGS as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \mu \mathbf{v}_{k+1} \tag{5}$$

$$\mathbf{v}_{k+1} = \mu \mathbf{v}_k - \eta_k \mathbf{H}_k^{\mathrm{NAQ}} \nabla f(\boldsymbol{\theta}_k + \mu \mathbf{v}_k), \tag{6}$$

$$\mathbf{H}_{k+1}^{\mathrm{NAQ}} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^{\mathrm{T}}) \mathbf{H}_k^{\mathrm{NAQ}} (\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^{\mathrm{T}}) + \rho_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}}, \tag{7}$$

$$\mathbf{s}_k = \boldsymbol{\theta}_{k+1} - (\boldsymbol{\theta}_k + \mu \mathbf{v}_k), \ \mathbf{y}_k = \nabla f(\boldsymbol{\theta}_{k+1}) - \nabla f(\boldsymbol{\theta}_k + \mu \mathbf{v}_k). \tag{8}$$

where μ is the momentum parameter and $\nabla f(\boldsymbol{\theta}_k + \mu \mathbf{v}_k)$ is the Nesterov's accelerated gradient. MoQ (Mahboubi et al. 2021) approximated $\nabla f(\boldsymbol{\theta}_k + \mu \mathbf{v}_k)$ in NAQ as a linear combination of past gradients. The acceleration of second order methods pave promising scope to numerous applications and is the focus of this research.

Research Overview

Training DNNs poses several challenges such as ill-conditioning, hyperparameter tuning, overparmeterization etc. Optimization in DNN, CNN, RNNs, and deep reinforcement learning (DRL), each encounter different difficulties and challenges based on the problem considered. For

example, RNNs popularly used in NLP, are powerful sequence models. But despite their capabilities in modeling sequences, RNNs are particularly very difficult to train long sequences with long-term dependencies due to the vanishing and/or exploding gradient problem. Hence several algorithms and architectures have been proposed to address the issues involved in training RNNs. Similarly, training NNs in reinforcement learning tasks is usually slow and challenging due to the training data being temporally correlated, non-stationary and presented as a continuous stream of experiences rather than batches like in supervised learning training, that makes it more prone to unlearning effective features over time. On the architecture front too, studies on overparameterization (Arora, Cohen, and Hazan 2018) suggest that increasing the depth of the NN architecture leads to faster training as adding layers increases expressive power, indicating that it is not an outcome of the so-called robust optimizer itself. Further, overparameterization coupled with large scale optimization and immense amount of data that needs to be processed in training a large neural network can in turn increase the computation and storage cost.

Thus this thesis aims to investigate if momentum accelerated second order methods such as NAQ and MoQ outperform conventional methods and avoid overparameterization, and more importantly, if they are robust, efficient and practical. To this end, we study the efficiency, robustness and accuracy of the Nesterov's accelerated quasi-Newton (NAQ) based optimization in the following cases: (1) deep neural network (full batch and stochastic) (2) time-series sequence modelling and (3) deep reinforcement learning, and develop practical algorithms for real-world problems.

Anticipated Research Contribution

- Study the behaviour of first order and second order quasi-Newton methods on different NN problems.
- Investigate and propose NAQ based optimization as a solution to these problems.
- Demonstrate robustness and efficiency of NAQ.
- Analyze computational cost and convergence.
- Investigate feasibility of Nesterov's acceleration to other algorithms in the quasi-Newton family.

Research Progress In this research, we began investigating the performance of NAQ in solving highly non linear problems (Indrapriyadarsini et al. 2018, 2020b) with the incorporation of a global convergence term. To facilitate large scale stochastic optimization, we proposed a stochastic NAQ variant and confirmed its robustness on feedforward NNs and CNNs along with a brief discussion on the computational cost both in full and limited memory forms (Indrapriyadarsini et al. 2019). To tackle the long sequence modelling issue in training RNNs, we proposed a stochastic NAQ variant with heuristic control that confirms the robustness of the proposed method in training RNNs compared to popular first order methods (Indrapriyadarsini et al. 2020a). The computation cost is also shown to be in the order of $O(d)$ and thus comparable with first order methods. In (Indrapriyadarsini et al. 2021), we extended the study to deep reinforcement learning (DRL) for solving global routing, a combinatorial optimization problem, and confirmed the robustness of our proposed method in the DRL case as well.

From the results obtained on the different classes of problems, we could confirm the practical robustness, accuracy and efficiency of NAQ. Our next set of goals towards this thesis include: (1) work on limited and stochastic variants of MoQ as a means of reducing the computation cost of oNAQ, (2) study the feasibility of the Nesterov's acceleration to other methods of the quasi-Newton family like SR1, and (3) provide theoretical analysis on convergence to further support the claims of our methods proposed thus far.

References

Arora, S.; Cohen, N.; and Hazan, E. 2018. On the optimization of deep networks: Implicit acceleration by overparameterization. In *ICML*, 244–253. PMLR.

Berahas, A. S.; Jahani, M.; Richtárik, P.; and Takáč, M. 2019. Quasi-newton methods for deep learning: Forget the past, just sample. *arXiv preprint arXiv:1901.09997*.

Indrapriyadarsini, S.; Mahboubi, S.; Ninomiya, H.; and Asai, H. 2018. Implementation of a Modified Nesterov's Accelerated Quasi-Newton Method on Tensorflow. In *IEEE ICMLA*, 1147–1154. IEEE.

Indrapriyadarsini, S.; Mahboubi, S.; Ninomiya, H.; and Asai, H. 2019. A Stochastic Quasi-Newton Method with Nesterov's Accelerated Gradient. In *ECML-PKDD*. Springer.

Indrapriyadarsini, S.; Mahboubi, S.; Ninomiya, H.; and Asai, H. 2020a. An adaptive stochastic Nesterov's Accelerated quasi-Newton method for training RNNs. *Nonlinear Theory and Its Applications, IEICE*, 11(4): 409–421.

Indrapriyadarsini, S.; Mahboubi, S.; Ninomiya, H.; Kamio, T.; and Asai, H. 2020b. A Neural Network Approach to Analog Circuit Design Optimization using Nesterov's Accelerated Quasi-Newton Method. In *IEEE ISCAS*, 1–1. IEEE.

Indrapriyadarsini, S.; Mahboubi, S.; Ninomiya, H.; Takeshi, K.; and Asai, H. 2021. A Nesterov's accelerated quasi-Newton method for global routing using deep reinforcement learning. *Nonlinear Theory and Its Applications, IEICE*, 12(3): 323–335.

Mahboubi, S.; Indrapriyadarsini, S.; Ninomiya, H.; Asai, H.; et al. 2021. Momentum acceleration of quasi-Newton based optimization technique for neural network training. *Nonlinear Theory and Its Applications, IEICE*, 12(3): 554–574.

Mokhtari, A.; and Ribeiro, A. 2015. Global convergence of online limited memory BFGS. *JMLR*, 16(1): 3151–3181.

Nesterov, Y. E. 1983. A method for solving the convex programming problem with convergence rate O(1/kˆ2). In *Dokl. akad. nauk Sssr*, volume 269, 543–547.

Ninomiya, H. 2017. A novel quasi-Newton-based optimization for neural network training incorporating Nesterov's accelerated gradient. *Nonlinear Theory and Its Applications, IEICE*, 8(4): 289–301.

Nocedal, J.; and Wright, S. J. 2006. *Numerical Optimization*. Springer Series in Operations Research. Springer, second edition.