Proposal For A Master Thesis

Zhiwei Liu

May 19, 2019

Affiliation: Master Thesis @ TU Berlin Faculty IV: EE & CS / DAI-Labor

Title: Parameter-free Online Optimization For Non-convex Problems

Candidate: Liu, Zhiwei

StudentID: 387571

Advisor: Shao, Weijia

Supervisor: Prof. Dr. Dr. h.c. Sahin Albayrak

Degree: Master of Science

Duration: June 1, 2019 - November 30, 2019 (6 months)

1 Introduction

1.1 Background

Over the last few years, machine learning technique, especially deep learning technique, has grown fast and powered many aspects of modern society [1]. Amazon.com shows personalized recommendation for distinct users based on its recommender system [2]. Google assistant helps owners make phone calls using its powerful speech and NLP technique. Thanks to computer vision, people are able to unlock their smart phones through facial recognition. The common method behind those applications is deep learning or its variants.

There are several reasons for such enormous progress. First of all, the improvement of computation power, especially GPU computing, makes training deep neural networks possible. Neural network methods have been proposed in 1980s (e.g., back-propagation [3], handwritten zip code recognition using MLP [4]). However, due to the limitation of computation power, those approaches could only be applied to some simple scenarios (i.e., small datasets and elementary neural networks). For decades, the computation power, GPU in particular, has gained a remarkable progress. Compared with CPU, GPU has larger bandwidth and more cores inside, which means it is good at processing computation operations which consume large memory blocks and has better concurrency performance. Those characteristics determine GPU is more suitable than CPU in deep learning area, since most of operations in neural network play with large matrices.

Secondly, state-of-the-art models and algorithms in recent years provide better generalization capacity to the real data from nature world comparing to traditional machine learning models. Not surprisingly, most of those new models are in forms of deep neural network with different architecture. Two typical cases are AlexNet and ResNet. AlexNet consists of eight layers: the

first five are convolutional layers, some of them followed by max-pooling layers, and the last three are fully connected layers[5]. ResNet [6] launches a skip-connection mechanism which skip over layers and performs very well. All those models are extremely complex and hard to train but receive a great success.

Meanwhile, the appearance of deep learning frameworks (e.g., TensorFlow[7] by Google, PyTorch[8] by Facebook, Keras[9], etc.) simplify and accelerate the design and development of complex neural networks. Also, the huge data collected data collected from Internet play an important role during the training and optimization of those large networks. For instance, ImageNet is a famous image database in computer vision area, which on average contains 500 images per node[10]. Such complete database becomes a useful resource for researchers.

1.2 Motivation

While sophisticated models and large datasets improve the generalization ability, effective optimizations on those models become a new emerging task needed to solve. However, most of those powerful models are in forms of non-convex functions which are extremely complicated. [11] shows that for a non-convex function, even verifying a feasible point is a local minimizer is NP-hard.

A common practice to solve such problems is to apply convex optimization theory and algorithms onto the non-convex problems. A typical example is the gradient descent (GD) algorithm which keeps approaching the point where the gradient is zero. For convex functions, such method must stop at the global optimum. For non-convex problems, such algorithm stops at saddle points (a.k.a. stationary points). To escape the saddle points, there are several intuitive ways, e.g., randomized choice of initial points[12], grid search on hyper-parameters, intermittent perturbations[13], etc.

While above solutions perform well in practice, there are still some problems to solve. One of the main problems is the existence of hyper-parameters from algorithms. Tuning (or searching) hyper-parameters is of inefficiency, especially when models are complex. Is there a way to eliminate those hyper-parameters? We also wonder, under such premise, is there an approach to escape from the saddle points efficiently?

2 Goal and Scope

In this thesis, we design and develop a parameter-free online algorithm to optimize non-convex problems which is able to escape from saddle points efficiently without tuning hyper-parameters manually. The aim of this work is expected to perform better (at least same) than current state-of-the-art algorithms under different measurements (e.g., loss on test set, convergence rate, etc.).

3 Related Work

There have been a number of papers working on this kind of problem. [14] provides a new abstraction for parameter-free optimization problems in the perspective of coin betting. It gives good performance but expensive time complexity and needs the bound of gradients. [15] proves the optimal regret bound while no prior information given. [16] uses Black-Box reductions to reduce the origin optimization problem to 2 sub problems, which reduces the parameter space into a 1D space and a unit ball space respectively. However, the diameter of parameters is still needed in this algorithm. Meanwhile, [17] proposes a graduated optimization strategy for

non-convex problems. While guaranteeing the global optimal for a non-convex problem, it needs several hyper parameters to tune.

4 Methodology

To achieve this goal, we mainly focus our research on the graduated optimization (a.k.a. continuation [18]) and reduction technique. The former one allows us to optimize a non-convex function from its roughly approximated convex function, which is meaningful to escape from saddle points. Meanwhile, the reduction technique provides a way to reduce the dimension of parameters. We explore and find an approach to combine those 2 technique and develop our new parameter-free algorithm framework.

5 Evaluation

To evaluate our algorithms, we concentrate on both theoretical part and empirical part. At first, We give a theoretical analysis to prove that the convergence rate of our algorithms are at least not worse compared with other state-of-the-art algorithms. Besides that, we conduct experiments in different types of applications (e.g., DNN, CNN, time series, etc.). The experimental results are used to support our theory analysis as well.

6 Outline

This thesis is separated into 6 chapters. Below are the short descriptions for each chapter.

Chapter 1 Introduction firstly introduces the background regarding the non-convex optimization problems and explains the motivation of parameter-free optimization. After that, it lists the objectives and declares the scope of this project. At the end, it gives a general overview for each chapter of this thesis.

Chapter 2 Preliminary lists notations and declares their usage regulations to avoid the abuse and confusion of notations and symbols. Later, it gives the necessary definitions that are frequently used in area of the optimization theory. Lastly, it introduces several core concepts which are the basis of this thesis.

Chapter 3 Related Work firstly introduces the convex optimization problems, which is the fundamental theory of this thesis. After that, some typical convex optimization algorithms (e.g., SGD, todo) are elaborated. Following this, a more complicated topic, non-convex optimization problem, is discussed. Also, related state-of-the-art algorithms are introduced and analyzed in this chapter.

Chapter 4 Algorithm explains our novel algorithm on non-convex optimization problems. It starts from showing the general framework of the algorithm. Following this, concrete details of each submodules are elaborated. After that, we give a theoretical analysis it and compare it with other state-of-the-art algorithms mentioned in the chapter 3 from the perspective of convergence rate.

Chapter 5 Evaluation is the evaluation part. It firstly introduces the experimental environments, which consists of the different datasets, platforms and applications. More specifically, it gives detailed explanations about the criteria to choose a dataset and the settings of different applications. Also, it tells the different measurements for the experimental results. Based on those measurements, the evaluation results are shown and explained. At the end, it comments on the usability and scalability of our novel algorithms.

Chapter 6 Conclusion mainly summarizes the thesis and gives a conclusion of this work. Meanwhile, it describes the problems that occurred during the work and shows how they are solved. Lastly, it gives an outlook about future work.

7 Time Plan

June 2019	• Read related literature • Design the algorithm framework • Write the *Introduction* and *Preliminary* chapter
July-Aug. 2019	• Implement algorithms (including our novel algorithm and some state-of-the-art algorithms that will be used for benchmarking) • Experiments: functionality test • Write the *Related Work* and *Algorithm* chapter
Sep.-Oct. 2019	• Further optimization and improvement • Experiments: benchmark with state-of-the-art algorithms • Write the *Evaluation* chapter and the *Abstract*
November 2019	• Write the *Conclusion* chapter • Finalize the *Abstract* and other minor issues • Prepare for the defense

Table 1: Task Schedule

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *NATURE*, vol. 521, pp. 436–444, 2015.

[2] R. Burke, A. Felfernig, and M. H. Göker, "Recommender Systems: An Overview," *AI Magazine*, vol. 32, p. 13, Jun 2011.

[3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct 1986.

[4] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, pp. 541–551, Dec 1989.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," tech. rep.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec 2015.

[7] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[8] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[9] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[11] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Mathematical Programming*, vol. 39, pp. 117–129, Jun 1987.

[12] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient Descent Converges to Minimizers," Feb 2016.

[13] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition," Mar 2015.

[14] F. Orabona and D. Pál, "Coin Betting and Parameter-Free Online Learning," Feb 2016.

[15] A. Cutkosky and K. Boahen, "Online Learning Without Prior Information," *Proceedings of Machine Learning Research*, vol. 65, pp. 1–35, 2017.

[16] A. Cutkosky and F. Orabona, "Black-Box Reductions for Parameter-free Online Learning in Banach Spaces," Feb 2018.

[17] E. Hazan, K. Y. Levy, and S. Shalev-Shwartz, "On Graduated Optimization for Stochastic Non-Convex Problems," Mar 2015.

[18] A. Blake and A. Zisserman, *Visual Reconstruction.* MIT Press, 1987.