Non-convex Optimization Theory for Deep Neural Networks

Fu Zhizheng

Nanjing University Suzhou Campus
1520 Taihu Road, Suzhou
Jiangsu Province, China, 210093
221900336@smail.nju.edu.cn

Abstract

With the rapid development of deep neural network, data analysis and other field of machine learning, more and more optimization problems generated. Most of them are non-convex optimization problems. In this article, I mainly discuss the significance of non-convex optimization problems in deep neural networks and summarize several important development of non-convex optimization in the past ten years.

Introduction

We have developed an in-depth understanding of convex optimization problems and devised a set of methods for their resolution, such as stochastic gradient descent, mini-batching, momentum, and variance reduction. However, in the majority of cases, particularly in the realm of deep neural networks, the problems we confront are inherently non-convex.

Taking deep neural networks as an example, while the activation function for each layer is convex, the composition of multiple hidden layers results in a non-convex objective function. General non-convex optimization is considered at least NP-hard due to unfavorable properties inherent to such problems, including the potential existence of numerous local minima, the presence of saddle points, expansive flat regions, and widely varying curvature.

In reality, there cannot be a universal algorithm to solve non-convex optimization problems in all cases, given the complexity and high dimensionality involved. Efficiently addressing these intricate non-convex problems holds significant importance for us.

Related works

There were many overview papers on non-convex optimization theory and its contribution in deep neural networks. They described their ideas and methods in great ways as well as their applications and directions for future research. Here, I am going to brief some outstanding overview papers on non-convex optimization theory for deep neural networks.

(Bottou, Curtis, and Nocedal 2018) presents a comprehensive theory of simple and general SG algorithms, discusses their practical behavior, and highlights opportunities for designing algorithms with improved performance which led to discussions about next-generation optimization methods for large-scale machine learning, including research into two main research directions: techniques to reduce noise in random directions and methods that exploit second-derivative approximations.

(**?**) provides an overview of optimization algorithms and theory for training neural networks. It is worth mentioning that this article focuses mainly on deep learning, and discusses in depth the various problems that may be encountered in the construction of deep neural networks. It focuses on SGD, Momentum and accelerated SGD and some Adaptive gradient methods.

(Sun et al. 2020) is a great review about findings and results on the global landscape of neural networks. It point out that wide neural nets may have sub-optimal local minima under certain assumptions and discuss visualization and empirical explorations of the landscape for practical neural nets.

(Lucas 2022) consisted of three main parts. The first part, they presented new optimization algorithms for deep learning, including Aggregated Momentum (AggMo) optimizer, Lookahead optimizer which both provide improved optimization performance for training deep neural network. The second part presented a bottom-up analysis of neural network loss landscapes. In the final part, the author adopted a top-down approach: the Monotonic Linear Interpolation property states that if a neural network is randomly initialized and trained to convergence then the loss on the line connection the initialization to the final solution will decrease monotonically.

Recent Advances

In recent years, there have been significant advancements in the field of non-convex optimization theory for deep neural networks. Researchers have explored various techniques and approaches to address the challenges posed by non-convexity in the objective functions of deep learning models. Here are some notable recent advances:

Landscape Analysis and Geometry

Understanding the geometric properties of the loss landscape associated with deep neural networks has become a

focal point. Recent studies have delved into the geometry of non-convex optimization landscapes, shedding light on the distribution of critical points, saddle points, and flat regions. This knowledge aids in developing optimization algorithms that navigate the landscape more effectively.

The significance of investigating the geometric properties of the loss landscape lies in addressing several key challenges in deep learning optimization: optimization difficulty, avoidance of local minima, escape from saddle points, accelerated convergence, algorithmic design foundation.

Hessian analysis Some regions of the loss landscape are incredibly ill-conditioned and difficult to traverse. While other regions can be closely approximated by well-conditioned quadratic objectives. Investigating the Hessian matrix, both theoretically and empirically, provides valuable insight into the loss landscape geometry and subsequent optimization behaviour of deep learning models (Lucas 2022). For example, (Kunin et al. 2021) showed that symmetry arising due to invariance in neural network architectures impose constraints on the Hessian matrix that lead to conservation laws under gradient flow.

Dynamical systems Also, neural network optimization is a dynamical system. Thus, we can utilize tools from dynamical systems analysis to better understand the loss landscape geometry(Lucas 2022). A more recent example given by (Tanaka and Kunin 2021) derived properties of optimization dynamics for deep neural networks which develop a theoretical framework to study the "geometry of learning dynamics" in neural networks and reveal a key mechanism of explicit symmetry breaking behind the efficiency and stability of modern neural networks.

Stochastic Gradient Descent (SGD) Variants

The large dimension of the decision variable in such problems motivates the use of first-order methods, which possess a cheap iteration. Moreover, the large amount of data motivates to use randomized methods such as stochastic gradient descent, which does not require to look through the whole dataset to make one step of the optimization procedure, thus making the iteration even cheaper(Danilova et al. 2021).

Building on the foundation of SGD, researchers have proposed and analyzed various variants to enhance optimization performance. Techniques such as adaptive learning rates, advanced momentum methods, and strategies to reduce noise in stochastic gradients have been explored. These advancements aim to mitigate challenges associated with convergence to sub-optimal solutions.

SGD with Large Step Sizes Learns Sparse Features (Andriushchenko et al. 2023) shows the features of the dynamics of the Stochastic Gradient Descent and presents empirical observations that commonly used large step sizes may lead the iterates to jump from one side of a valley to the other causing loss stabilization, and this stabilization induces a hidden stochastic dynamics that biases it implicitly toward simple predictors.

Adaptive gradient methods: AdaGrad, RMSProp, Adama and more Make Adama as an example, Adama is the combination of RMSProp and the momentum method, which is well suited for problems that are large in terms of data and/or parameters. The method is also appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients(Kingma and Ba 2017).

Second-Order Optimization Methods

While first-order methods like SGD dominate the training of deep neural networks, there is a growing interest in second-order optimization methods. These methods leverage second-order information, such as Hessians, to achieve faster convergence and more accurate updates. Adapting these methods to the non-convex setting is an active area of research.

Second-Order Optimization for Non-Convex Machine Learning (Xu, Roosta-Khorasani, and Mahoney 2018) is an opening piece. In this study, they conduct extensive empirical evaluations on a category of Newton-type methods, specifically sub-sampled variants of trust region (TR) and adaptive regularization with cubics (ARC) algorithms, applied to non-convex machine learning problems. Their findings reveal that these methods not only exhibit computational competitiveness when compared to manually tuned stochastic gradient descent (SGD) with momentum, achieving comparable or superior generalization performance, but also demonstrate high robustness across various hyperparameter settings. Additionally, they highlight a distinctive advantage of these Newton-type methods over SGD with momentum—their effective utilization of curvature information enables smooth navigation through flat regions and avoidance of saddle points.

However, it's crucial to acknowledge the challenges associated with the adoption of second-order optimization methods. The computation and storage requirements for calculating and storing Hessian matrices can be computationally expensive, especially for large-scale deep neural networks. Addressing these computational challenges remains an ongoing area of research as researchers strive to make second-order optimization methods more scalable and applicable to real-world deep learning tasks.

Regularized Newton Method with Global $O(1/k^2)$ Convergence In this article, the author present a Newton-type method that converges fast from any initialization and for arbitrary convex objectives with Lipschitz Hessians. The iterates are given by $x^{k+1} = x^k - (\nabla^2 f(x^k) + \sqrt{H||\nabla f(x^K)||}I)^{-1}\nabla f(x^k), H > 0$ The method is the first variant of Newton's method that has both cheap iterations and provably fast global convergence(Mishchenko 2023).

Summary

The article explores the significance of non-convex optimization problems in the context of deep neural networks. Despite the development of methods for convex optimization, non-convex problems dominate in deep learning. The challenges associated with non-convex optimization, such as

multiple local minima and saddle points, make it a complex and high-dimensional problem.

The related works section discusses various overview papers on non-convex optimization theory for deep neural networks. These papers delve into optimization algorithms, practical behaviors, and opportunities for improvement in large-scale machine learning. Additionally, they explore the global landscape of neural networks and present new optimization algorithms like Aggregated Momentum and Lookahead.

In recent years, significant advances have been made in non-convex optimization theory for deep neural networks. The landscape analysis and geometry section emphasizes the importance of understanding the geometric properties of loss landscapes, including Hessian analysis and dynamical systems. Stochastic Gradient Descent (SGD) variants, such as adaptive gradient methods, are discussed, highlighting their effectiveness in handling large datasets.

The second-order optimization methods section introduces the growing interest in leveraging second-order information, like Hessians, for faster convergence. While acknowledging the challenges, such as computational expenses for large-scale networks, the article presents studies on regularized Newton methods with global convergence.

Prospects

These recent advances collectively contribute to a deeper understanding of non-convex optimization challenges in deep neural networks and pave the way for more efficient training procedures, improved model architectures, and enhanced generalization capabilities. Ongoing research continues to explore innovative solutions to further push the boundaries of deep learning optimization.

Future research may focus on refining and developing new optimization algorithms, addressing challenges related to loss landscape geometry, and improving the scalability of second-order optimization methods. Additionally, advancements in understanding and mitigating issues like saddle points and convergence to sub-optimal solutions are areas for exploration.

Exploring the intersection of optimization and interpretability in deep learning models could provide valuable insights. Continued efforts to bridge the gap between theory and practical applications will contribute to the efficient solution of complex and high-dimensional non-convex optimization problems in the field of deep neural networks.

References

Andriushchenko, M.; Varre, A. V.; Pillaud-Vivien, L.; and Flammarion, N. 2023. SGD with Large Step Sizes Learns Sparse Features. 202: 903–925.

Bottou, L.; Curtis, F. E.; and Nocedal, J. 2018. Optimization Methods for Large-Scale Machine Learning.

Danilova, M.; Dvurechensky, P.; Gasnikov, A.; Gorbunov, E.; Guminov, S.; Kamzolov, D.; and Shibaev, I. 2021. Recent Theoretical Advances in Non-Convex Optimization.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization.

Kunin, D.; Sagastuy-Brena, J.; Ganguli, S.; Yamins, D. L. K.; and Tanaka, H. 2021. Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics.

Lucas, J. 2022. Optimization and loss landscape geometry of deep learning.

Mishchenko, K. 2023. Regularized Newton Method with Global $\mathcal{O}(1/k^2)$ Convergence. *SIAM Journal on Optimization*, 33(3): 1440–1462.

Sun, R.; Li, D.; Liang, S.; Ding, T.; and Srikant, R. 2020. The Global Landscape of Neural Networks: An Overview. *IEEE Signal Processing Magazine*, 37(5): 95–108.

Tanaka, H.; and Kunin, D. 2021. Noether's Learning Dynamics: Role of Symmetry Breaking in Neural Networks.

Xu, P.; Roosta-Khorasani, F.; and Mahoney, M. W. 2018. Second-Order Optimization for Non-Convex Machine Learning: An Empirical Study. arXiv:1708.07827.