An Attempt to Escape the Deep Saddle Points

Ankit Goyal
Department of Electrical Engineering
Indian Institute of Technology Kanpur
Kanpur, India
ankgoyal@iitk.ac.in

Md Enayat Ullah
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Kanpur, India
enayat@iitk.ac.in

Abstract

Stochastic Gradient Descent (SGD) is a popular convex optimization technique, which has been widely applied on various non-convex optimization problems, including the ones encountered while training Deep Neural Networks (DNN). In spite of their empirical success, there exists limited understanding about the theoretical applicability of SGD in non-convex settings. Recent works have shown that the crucial challenge in high-dimensional non-convex optimization problems is averting convergence at saddle points and spurious local minima. We propose to build upon the recent results, which guarantee SGD to escape saddle points, if the optimization problem exhibits the strict-saddle property. In particular, our intention is to extend the analysis for Tensor Decomposition to two-layered Neural Networks (NN), as both these non-convex problems suffer from escalation of saddle points in high-dimensional settings, owing to their symmetry. We further aim to expand the analysis of the classical two-layered NN to encompass the deep learning paradigm in general.

1 Introduction

The idea of biologically inspired artificial neural networks existed since the 1980's [1], however their true potential was realized with the advent of computational power and large-scale data. Recently, Deep Neural Networks (DNN) have emerged as a hammer cracking down a plethora of machine learning problems [2, 3]. But despite their empirical success, there have been limited studies concerning their theoretical properties.

In general, the objective function for training a DNN is non-convex. However, initially researchers used off-the-shelf convex optimization techniques like Stochastic Gradient Descent (SGD) and L-BFGS [4], for training DNNs. In particular, SGD has been extensively studied in convex settings [5], but its effectiveness in non-convex settings is still not completely understood. Dauphin et al. in [6] showed that for non-convex optimization in high-dimensional settings, it is the proliferation of saddle-points, and not local minima that pose problem for gradient descent based methods. They also argued that in such high-dimensional settings, it is sufficient to converge to a local minimum, as the value of objective function at a local minimum would be very close to its value at the global minimum. Following these lines, they suggested a gradient descent based method for non-convex optimization, that uses second-order Hessian information, to avoid avoid/escape the saddle point. Furthermore, Ge et al. in [7] identified a strict-saddle property which allowed them to use SGD for escaping saddle points even without using the second-order information, and thereby avoiding the computational and memory overhead. They illustrated the effectiveness of this approach for Tensor Decomposition, which is a widely used non-convex optimization problem. They showed how objective function of Tensor Decomposition can be reformulated so as to obey the strict-saddle property, and thereby proposed a framework for Online Tensor Decomposition.

There have always been attempts to study non-convex problems, by identifying properties similar to the convex ones. Recent works have analyzed non-convex optimization problems with added constraints like Restricted Strong Convexity (RSC) [8]. However, since these constraints ensure that there is a unique stationary point [7], the analysis of these problems cannot be transferred to high-dimensional DNN cost function optimization, as here we have a mushrooming

of saddle points. Tensor Decomposition problem exhibits a symmetry in its solution, whereby if $(v_1, v_2, ..., v_d)$ is a solution, then for any permutation π and any sign flips $\kappa \in \{\pm 1\}^d$, $(.., \kappa_i v_{\pi(i)}, ...)$ is also a valid solution. Such a symmetry results in escalation of saddle points. Neural Networks (NN) exhibit a similar symmetry in its weight space, and thereby suffers from an excess of saddle points in its objective function. Saxe et al. in [9] show how scaling symmetries in Deep linear MLP give rise to saddle points. At the same time, DNN architecture serve as a natural extension to simple NN. The above factors serve as motivation for analyzing NN using the framework proposed by Ga et al. [7]. We propose to reformulate the objective function for training NN, so that it falls in the paradigm of strict-saddle functions. We further plan to analyze the generalisability of such a formulation for optimizing a DNN architecture.

2 Problem Definition

Note: This sections draws heavily from [7]. We have used the notations presented by Ge et al. to present the mathematical framework for our proposal.

2.1 Mathematical Preliminaries

Stochastic Gradient Descent (SGD)

Stochastic gradient descent is a gradient descent based optimization routine which solves the following problem:

$$w = \arg\min_{w \in \mathbb{R}^d} f(w), \text{ where } f(w) = \mathbb{E}_{x \sim \mathcal{D}}[\phi(w, x)] \tag{1}$$

The data point x is drawn from some unknown distribution $\mathcal{D}$, and ϕ is a loss function that is defined for a pair (x, w). The aim is to minimize the expected loss $\mathbb{E}[\phi(w, x)]$. The parameter updation takes place following a stochastic gradient

$$w_{t+1} = w_t - \eta \nabla_{w_t} \phi(w_t, x_t), \tag{2}$$

where x_t is a random sample drawn from distribution $\mathcal{D}$ and η is the learning rate.

Strict-Saddle Property

Definition 1. A twice differentiable function $f(w)$ is strict-saddle, if all its local minima have $\nabla^2 f(w) \succ 0$ and all its other stationary points satisfy $\lambda_{min}(\nabla^2 f(w)) < 0$.

Definition 2. A twice differentiable function $f(w)$ is $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, if for any point w at least one of the following is true

1. $\|\nabla f(w)\| \geq \epsilon$.
2. $\lambda_{min}(\nabla^2 f(w)) \leq -\gamma$.
3. There is a local minimum $w^\star$ such that $\|w - w^\star\| \leq \delta$, and the function $f(w')$ restricted to 2δ neighborhood of $w^\star$ ($\|w' - w^\star\| \leq 2\delta$) is α-strongly convex.

Tensor Decomposition

A tensor is a natural generalization of scalars and vectors to arbitrary dimensions. Tensors can be constructed from tensor products. $(u \otimes v)$ denotes a second order tensor where $(u \otimes v)_{i,j} = u_i v_j$. This generalizes to higher order and $u^{\otimes 4}$ denotes the 4-th order tensor

$$[u^{\otimes 4}]_{i_1, i_2, i_3, i_4} = u_{i_1} u_{i_2} u_{i_3} u_{i_4}.$$

A 4-th order tensor $T \in \mathbb{R}^{d^4}$ has an orthogonal decomposition if it can be written as

$$T = \sum_{i=1}^{d} a_i^{\otimes 4}, \tag{3}$$

where a_i's are orthonormal vectors that satisfy $\|a_i\| = 1$ and $a_i^T a_j = 0$ for $i \neq j$. Vectors a_i's are called the components of this decomposition.

Tensor decomposition is a non-convex optimization problem where the objective function is generally formulated as follows:

$$\min_{\forall i, \|u_i\|^2=1} \quad \|T - \sum_{i=1}^{d} u_i^{\otimes 4}\|_F^2. \tag{4}$$

Ge et al. [7] proposed an alternative objective function because they were unable to prove strict-saddle property for the objective in Eq. 4. The following objective function for Tensor Decomposition was proved to satisfy strict-saddle property.

$$\min_{\forall i, \|u_i\|^2=1} \quad \sum_{i \neq j} T(u_i, u_i, u_j, u_j), \tag{5}$$

Neural Network(NN)

We intend to analyze a two-layer feed forward NN wherein each layer is fully connected with the adjacent layer as shown in Fig. 1. The input features are fed to the first layer. Each node in a neural net receives a set of inputs from other connected nodes, and outputs a value which is a linear/non-linear function applied on the weighted sum of the inputs. The weights are updated so as to minimize the empirical loss on the training data. The objective function of a generic NN is:

$$J(w) = \frac{1}{N} \sum_{n=1}^{N} L(\hat{y}_i, y_i) \tag{6}$$

where $\hat{y}_i$ is the predicted response, y_i is the ground truth, and L is the loss function.

2.2 Proposed Work

We plan to investigate the objective function of a two-layer NN and check if any of the popular ones satisfy the strict-saddle property. Our aim is to analyze and formulate an objective function for NN which suits the strict-saddle framework. We further aim to empirically validate the effectiveness of the modified objective against traditional approaches on some benchmark datasets. We plan to explore the generalizability of this analysis to deeper architectures such as CNN, RNN. We further intend to study other non-convex problems such as sparse recovery in this setting.

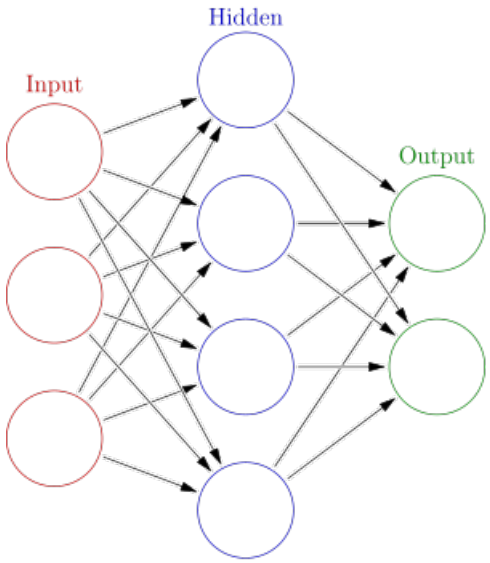

Figure 1: A two-layer Neural Network

Source: Wikipedia

References

[1] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[4] Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V Le, and Andrew Y Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 265–272, 2011.

[5] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[6] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

[7] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. *arXiv preprint arXiv:1503.02101*, 2015.

[8] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[9] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.