Spin Glass Models in Neural Networks

Abstract

Neural networks have been the focus of the most exciting research for over a decade in the field of machine learning. They have proven to be an incredibly powerful and flexible model that generalizes well even in an over-parameterised setting. However, as powerful as it is, we are still struggling to understand their inner-workings. This paper explores the advancements made in this endevour through the use of the spin glass model taken from statistical physics.

Introduction

While neural networks are incredibly powerful, they are very difficult for a human to interprate and many optimisation techniques still only come from empirical evidence instead of theoretical justification. The relationship between spin glass models and neural networks was recognised in the very early stages of the existance of neural networks [1]. While there has been progress in analyzing how DNNs work mathematically [2], our understanding remains incomplete. In this project I will discuss the involvement of the spin glass model in aiding explainability and optimisation in DNN. We can take advantage of the spin glass' non-convex nature to understand why deep neural networks (DNNs) are so easily optimizable dispite their also non-convex loss landscape while still using something as simple as gradient decent.

Background

While this approach to viewing, understanding and optimising neural networks is relatively niche, I still think it offers meaningful insight into the mechanics of such a system through the network's loss landscape. The loss landscape of a neural network is, put simply, a mapping of each possible weight configuration to a corresponding loss value. Naturally, it would be infeasible to, model and then test every single variation of weights to find the one that gives of the best performance, so we must find a method that can both find a minimum training loss with the hope of also having good generalisation, and also be effective at doing so. The problem is a loss landscape is non-convex and most of the time in very high dimensions, therefore making it a complex problem to solve. Yet, minimising the loss function has been empirically very tractable [3]. Simple algorithms such as stochastic gradient descent and Adam are commonly used which raises a key question: why is such a complex problem so easy to solve?

To answer this intriguing question we must look toward the physical spin glass magnet, a spin glass is a disordered magnetic system with randomly interacting spins, resulting in a rugged energy landscape with many local minima. This type of landscape is similar to that of a neural network's loss surface, which comes from the non-convex nature of objective functions (e.g. cross-entropy, mean squared error) and overparameterization. In particular, the p-spin (or H-spin as it is refered to in the literature) spherical spin glass model provides a useful analogy for understanding the geometry and critical points of these loss landscapes.

By modeling a neural network's loss landscape using a H-spin spherical spin glass, we can study the optimization process from a statistical physics perspective. Under assumptions such as parameter redundancy, independence of variables, and uniformity, the neural network's loss function becomes analogous to the Hamiltonian of a spin glass system [4].

Theoretical Findings

The earliest attempts to mathematically understand the optimistion problem of a non-convex loss function was [4], through the use of spin glass models. By modeling the loss landscape using a spin glass we can describe the landscape in terms of energy levels, from the ground state $E_0(H)$ (the global minimum) to the energy barrier $E_\infty(H)$, defined by:

$$E_\infty(H) = 2\sqrt{\frac{H-1}{H}}$$

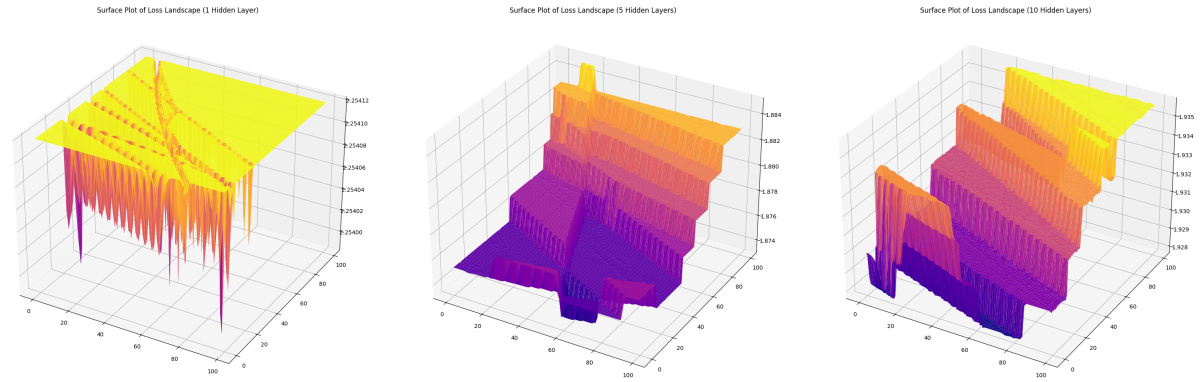

Figure 1: The difference in smoothness in loss landscapes between neural networks with different depths is very evident when comparing the network with one hidden layer (leftmost figure) and the networks with five and ten hidden layers (middle and rightmost figure respectively)

where H is the dimensionality of the system. Critical points with energy higher than $E_\infty(H)$ are exponentially likely to be high index saddle points [5]. Within the band $(-\Lambda E_0(H), -\Lambda E_\infty(H))$, where Λ is given in **Definition 3.1** of [4], the probability of the existence of a saddle point rapidly approaches zero as the dimensionality increases. This implies that local minima are more densely located in this region.

In practice, the goal is not to find the global minimum—doing so becomes exponentially difficult as network size increases and may lead to overfitting—but rather to identify a good local minimum that generalizes well. Spin glass theory provides an effective framework for understanding why stochastic gradient descent often converges to such solutions.

Spin glass theory has also enabled us to understand more about the role of network depth. Shallow neural networks have been shown to be just as effective as DNNs in their accuracy [6], however the differences between the two lies in their optimisability. Allowing depth to vary while fixing the number of parameters in place, we see that minima start to coagulate together and the trade-off between the width and depth of minima becomes weaker as the depth increases [2]. We can also see the number of critial points N, as described in **Equation 5** of [2]

$$N = \frac{(H-1)^\Lambda - 1}{H-2}$$

is non-monotonic, meaning at a certain depth, the number of critical points starts to decrease giving the minimisor less "bad" critical points to potentially land on instead of a good local minimum. Because of this, the fact that minima become more clustered with wider basins, deeper networks become much easier to optimise than shallower networks even though accuracy may not change with depth.

Experimental Findings

An experiment was carried out to confirm the theoretical proposal that increasing the depth of the network while keeping the number of parameters the same, eliminating "bad minima" and also increasing the width of minima's basin. For this experiment I used the code from [2] in order to visualise the landscape, using the MNIST dataset to train the model. For implemenation details, see [7]. As seen in Figure 1 the loss landscape formed with only one hidden layer is jagged with shallow and narrow minima, however when the depth of the neural network increases, the landscape becomes smoother and the minima become much broader. This aligns empirically with the theoretical findings of [3].

In Figure 2, we see the empirical evidence of the similarities between the p-spin spherical spin glass model's energy landscape and the loss landscape of neural networks. Using the sklearn linnerud dataset, we can see that it produces a complex lanscape for both the neural network (while it is very subtle there are two minima in

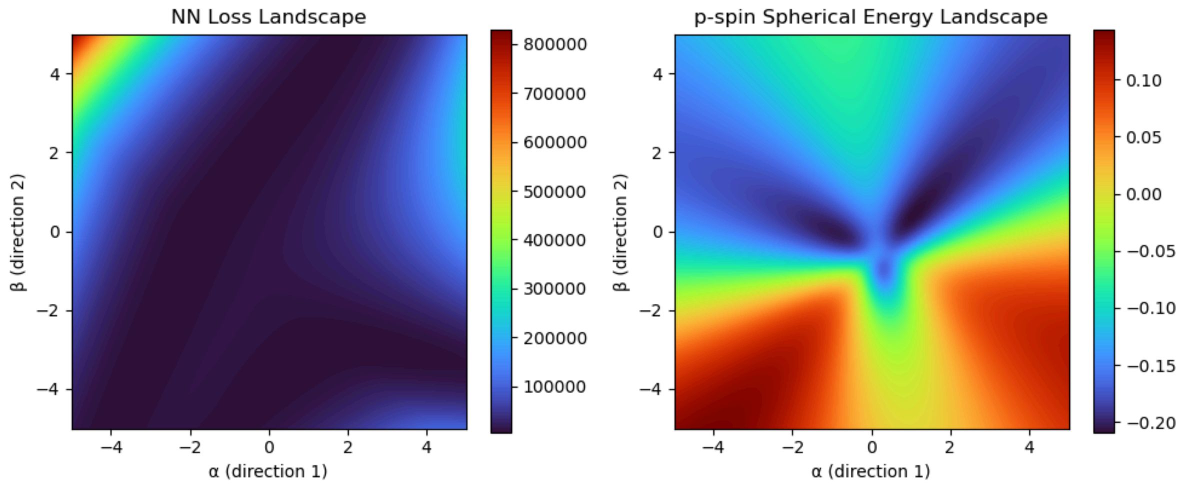

Figure 2: The similarity between the energy landscape of the spin glass model and the loss landscape of a neural network. We can see that they both have complex, non-convex landscapes which implies similar properties.

the loss landscape plot with a maxima splitting them through the middle) and the spin glass model. By proving that they both are non-convex we can also say that they share multiple other properties such as the minima band and the exponential growth of critical points as the system increases.

Conclusion

In summary I have explored how, through the application of spin glass theory, we can understand more about neural networks and their optimisability because of the similarities in the spin glass model's energy landscape and the neural network's loss landscape which can be seen in Figure 2 showing both of their non-convexities and complex landscapes. I have shown that, according to spin glass theory minima are confined within a band. Higher than this band, the frequency of minima decreases exponentially as the number of saddle points start to increase exponentially [4]. I also showed that, while initially the number of critical points increases when the depth of the network increases, it begins to decrease allowing the network to optimise easier as there is less saddle points for the optimiser to get stuck on. The minima also deepen and broaden as the network deepens which can be seen in Figure 1.

References

[1] D. J. Amit, H. Gutfreund, and H. Sompolinsky, “Spin-glass models of neural networks,” *Physical Review A*, vol. 32, no. 2, pp. 1007–1018, 1985, doi: 10.1103/PhysRevA.32.1007.

[2] S. Becker, Y. Zhang, and A. A. Lee, “Geometry of Energy Landscapes and the Optimizability of Deep Neural Networks,” *arXiv preprint arXiv:1808.00408*, 2018.

[3] H. Liao *et al.*, “Exploring Loss Landscapes through the Lens of Spin Glass Theory,” *arXiv preprint arXiv:2407.20724*, 2024.

[4] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, “The Loss Surfaces of Multilayer Networks,” *AISTATS*, 2015.

[5] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, “Identifying and attacking the saddle point problem in high-dimensional non-convex optimization,” *arXiv preprint arXiv:1406.2572*, 2014.

[6] L. J. Ba and R. Caruana, “Do Deep Nets Really Need to be Deep?,” *arXiv preprint arXiv:1312.6184*, 2014.

[7] M. D. Bernardi, “loss-landscapes.” 2020.