CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2023-0020

Vol. **5**, No. 2, pp. 350-389
June 2024

Embedding Principle in Depth for the Loss Landscape Analysis of Deep Neural Networks

Zhiwei Bai[1], Tao Luo[1,2], Zhi-Qin John Xu[1,*] and Yaoyu Zhang[1,3,*]

[1] *School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[2] *CMA-Shanghai, Shanghai Artificial Intelligence Laboratory, Shanghai 200240, P.R. China.*
[3] *Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 200240, P.R. China.*

Received 16 May 2023; Accepted 7 January 2024

Summary. In this work, we delve into the relationship between deep and shallow neural networks (NNs), focusing on the critical points of their loss landscapes. We discover an embedding principle in depth that loss landscape of an NN "contains" all critical points of the loss landscapes for shallower NNs. The key tool for our discovery is the critical lifting that maps any critical point of a network to critical manifolds of any deeper network while preserving the outputs. To investigate the practical implications of this principle, we conduct a series of numerical experiments. The results confirm that deep networks do encounter these lifted critical points during training, leading to similar training dynamics across varying network depths. We provide theoretical and empirical evidence that through the lifting operation, the lifted critical points exhibit increased degeneracy. This principle also provides insights into the optimization benefits of batch normalization and larger datasets, and enables practical applications like network layer pruning. Overall, our discovery of the embedding principle in depth uncovers the depth-wise hierarchical structure of deep learning loss landscape, which serves as a solid foundation for the further study about the role of depth for DNNs.

AMS subject classifications: 68T07
Key words: Deep learning, loss landscape, embedding principle.

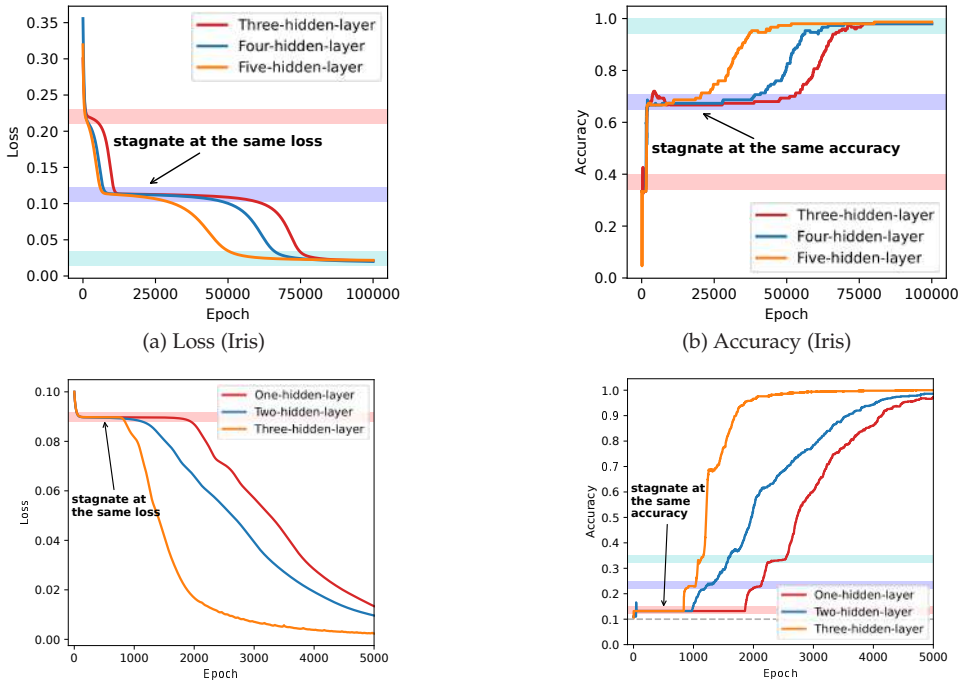
1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in various fields, such as computer vision [18], natural language processing [4], and numerous scientific computing applications [2,10,24]. Despite their widespread adoption and empirical achieve-

*Corresponding author. *Email addresses:* bai299@sjtu.edu.cn (Z. Bai), luotao41@sjtu.edu.cn (T. Luo), xuzhiqin@sjtu.edu.cn (Z. Xu), zhyy.sjtu@sjtu.edu.cn (Y. Zhang)

ments, our theoretical understanding of DNNs, particularly regarding their loss landscape and training dynamics, remains limited. The loss landscape of a DNN essentially characterizes the optimization problem encountered during the network's training process. The study of this landscape is of paramount importance as it directly influences not only the efficiency and final outcome of the training process, but also the generalization in overparametered case. Regrettably, the high-dimensionality and non-convex nature of DNNs render their loss landscapes notoriously challenging to comprehend and navigate. The recent discovery of the embedding principle [9,20,30,32] offers insights for analyzing the loss landscape of networks and establishes connections between the loss landscapes of neural networks with varying widths. However, considering the extreme importance of depth for DNNs, it prompts us to question whether a relationship exists between the loss landscapes of networks with different depths. In this paper, we strive to address this fundamental question by conducting a thorough analysis of critical points across varying network depths.

Our theoretical investigation is motivated by the following experimental observations, which hint at the existence of an embedding relationship in depth. As illustrated in Fig. 1, the training of NNs with varying hidden layers, learning the Iris and MNIST



(a) Loss (Iris) (b) Accuracy (Iris)

(c) Loss (MNIST) (d) Accuracy (MNIST)

Figure 1: The training dynamics of networks of different depths exhibit similarity. (a, c) The training loss for NNs of varying depths on the Iris and MNIST datasets, respectively. (b, d) The corresponding training accuracy for NNs of varying depths on the Iris and MNIST datasets, respectively. The color-coded areas indicate periods of slow change in training loss or training accuracy, indicating a possible encounter with a saddle point.