CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2023-0016

Vol. 5, No. 3, pp. 448-514
September 2024

Phase Diagram of Initial Condensation for Two-Layer Neural Networks

Zheng-An Chen[1,2], Yuqing Li[1,2,*], Tao Luo[1,2,3,4], Zhangchen Zhou[1,3] and Zhi-Qin John Xu[1,3,*]

[1] *School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[2] *CMA-Shanghai, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[3] *Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[4] *Shanghai Artificial Intelligence Laboratory, Shanghai 200232, P.R. China.*

Received 7 April 2023; Accepted 27 February 2024

Abstract. The phenomenon of distinct behaviors exhibited by neural networks under varying scales of initialization remains an enigma in deep learning research. In this paper, based on the earlier work [Luo *et al.*, J. Mach. Learn. Res., 22:1–47, 2021], we present a phase diagram of initial condensation for two-layer neural networks. Condensation is a phenomenon wherein the weight vectors of neural networks concentrate on isolated orientations during the training process, and it is a feature in non-linear learning process that enables neural networks to possess better generalization abilities. Our phase diagram serves to provide a comprehensive understanding of the dynamical regimes of neural networks and their dependence on the choice of hyperparameters related to initialization. Furthermore, we demonstrate in detail the underlying mechanisms by which small initialization leads to condensation at the initial training stage.

AMS subject classifications: 68U99, 90C26, 34A45

Key words: Two-layer neural network, phase diagram, dynamical regime, condensation.

1 Introduction

In deep learning, one intriguing observation is the distinct behaviors exhibited by neural networks (NNs) depending on the scale of initialization. Specifically, in a particular regime, NNs trained with gradient descent can be viewed as a kernel regression predictor

*Corresponding author. *Email addresses:* `zhengan_chen@sjtu.edu.cn` (Z. Chen), `liyuqing_551@sjtu.edu.cn` (Y. Li), `luotao41@sjtu.edu.cn` (T. Luo), `zczhou1115@sjtu.edu.cn` (Z. Zhou), `xuzhiqin@sjtu.edu.cn` (Z. J. Xu)

known as the neural tangent kernel (NTK) [5,10,11,15], and Chizat *et al.* [4] identify it as the lazy training regime in which the parameters of overparameterized NNs trained with gradient based methods hardly varies. However, under a different scaling, the gradient flow (GF) of NN shows highly nonlinear features and a mean-field analysis [3,19,23,24] has been established for infinitely wide two-layer networks to analyze its behavior. Additionally, small initialization is proven to give rise to condensation [16,18,30,31], a phenomenon where the weight vectors of NNs concentrate on isolated orientations during the training process. This is significant as NNs with condensed weight vectors are equivalent to “smaller” NNs with fewer parameters, as revealed by the embedding principle (the loss landscape of a deep neural network (DNN) “contains” all the critical points of all the narrower DNNs [28,29]), thus reducing the complexity of the output functions of NNs. As the generalization error can be bounded in terms of the complexity [1], NNs with condensed parameters tend to possess better generalization abilities. In addition, the study of the embedding principle also found the number of the descent directions in a condensed large network is no less than that of the equivalent small effective network, which may lead to easier training of a large network [28,29].

Taken together, identifying the regime of condensation and understanding the mechanism of condensation are important to understand the non-linear training of neural networks. Our contributions can be categorized into two aspects.

Firstly, we established the phase diagram of initial condensation for two-layer neural networks with a wide class of smooth activation functions, as illustrated Fig. 1. Note that the phase diagram drawn in [16] is only for two-layer wide ReLU networks and the phase

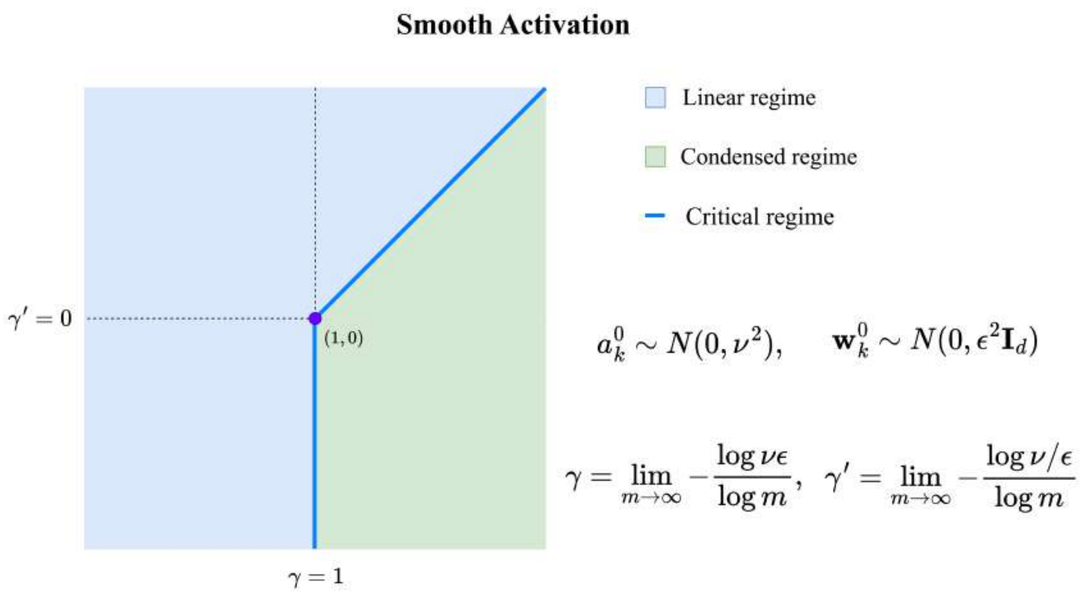

Figure 1: Phase diagram of two-layer NNs.