Journal of Computational Mathematics
Vol.44, No.2, 2026, 369–393.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2412-m2024-0083

LOSS SPIKE IN TRAINING NEURAL NETWORKS*

Xiaolong Li
*School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC,
Shanghai Jiao Tong University, Shanghai 200240, China*
Zhi-Qin John Xu[1)]
*School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC,
Shanghai Jiao Tong University, Shanghai 200240, China;
Key Laboratory of Marine Intelligent Equipment and System, Ministry of Education,
Shanghai 200240, China
Email: xuzhiqin@sjtu.edu.cn*
Zhongwang Zhang
*School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC,
Shanghai Jiao Tong University, Shanghai 200240, China*

Abstract

In this work, we investigate the mechanism underlying loss spikes observed during neural network training. When the training enters a region with a lower-loss-as-sharper structure, the training becomes unstable, and the loss exponentially increases once the loss landscape is too sharp, resulting in the rapid ascent of the loss spike. The training stabilizes when it finds a flat region. From a frequency perspective, we explain the rapid descent in loss as being primarily influenced by low-frequency components. We observe a deviation in the first eigendirection, which can be reasonably explained by the frequency principle, as low-frequency information is captured rapidly, leading to the rapid descent. Inspired by our analysis of loss spikes, we revisit the link between the maximum eigenvalue of the loss Hessian ($\lambda_{\max}$), flatness and generalization. We suggest that $\lambda_{\max}$ is a good measure of sharpness but not a good measure for generalization. Furthermore, we experimentally observe that loss spikes can facilitate condensation, causing input weights to evolve towards the same direction. And our experiments show that there is a correlation (similar trend) between $\lambda_{\max}$ and condensation. This observation may provide valuable insights for further theoretical research on the relationship between loss spikes, $\lambda_{\max}$, and generalization.

Mathematics subject classification: 65N38, 65N30.
Key words: Neural Network, Loss Spike, Frequency Principle, Maximum Eigenvalue, Flatness, Generalization, Condensation.

1. Introduction

Many experiments have observed a phenomenon, called the edge of stability (EoS) [4,8,10,43,57], that learning rate (η) and sharpness (i.e. the largest eigenvalue of Hessian) no longer behave as in traditional optimization, sharpness hovers at $2/\eta$ while the loss continues decreasing, albeit non-monotonically. Training with a larger learning rate leads to a solution with smaller $\lambda_{\max}$. Since $\lambda_{\max}$ is often used to indicate the sharpness of the loss landscape, a larger learning rate results in a flatter solution. Intuitively as shown in Fig. 1.1, the flat solution is more robust

* Received April 16, 2024 / Revised version received September 19, 2024 / Accepted December 2, 2024 / Published online January 13, 2025 /

[1)] Corresponding author

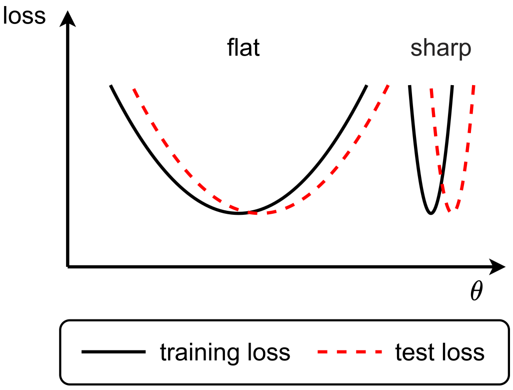

Fig. 1.1. Schematic illustration of an ideal explanation for why flat solutions generalize well [23].

to perturbation and has better generalization performance [19, 23]. Therefore, training with a larger learning rate would achieve better generalization performance. In this work, we argue this intuitive analysis in Fig. 1.1 with $\lambda_{\max}$ as the sharpness measure, which encounters difficulty in NNs through the study of loss spikes.

In a neural network training process, one may sometimes observe a phenomenon of loss spike (typical examples in Fig. 3.1), where the loss rapidly ascends and then descends to the value before the ascent. We show a special loss landscape structure underlying the loss spike, which is called a lower-loss-as-sharper (LLAS) structure. In the LLAS structure, the training is driven by descending the loss while entering an increasingly sharp region. Once the sharpness is too large, the loss would ascend exponentially fast. To explain why the loss can descend so fast, we provide a frequency perspective analysis. We find that the deviation in the ascending stage is dominated by low-frequency components. Based on the frequency principle [47, 48] that low-frequency converges faster than high-frequency, we rationalize the fast descent.

The study of loss spike provides an important information that the deviation at the first eigendirection is dominated by low-frequency. We then further argue the link between $\lambda_{\max}$ flatness and generalization. In real-world datasets, low-frequency information is often dominant and well-captured by both the training data and the test data. Therefore, the training can learn low-frequency well. Since the sharpest direction, indicated by the maximum eigenvalue of the loss Hessian, relates more to the low-frequency, a solution with good generalization and a solution with bad generalization have little difference in the sharpest direction, verified by a series of experiments. Hence, $\lambda_{\max}$ with the intuitive explanation in Fig. 1.1 encounters difficulty in understanding the generalization of neural networks, such as why a larger learning rate results in better generalization for networks with EoS training.

We also find that a loss spike can facilitate condensation, that is, the input weights of different neurons in the same layer evolve towards the same, which would reduce the network's effective size. Condensation is a non-linear feature learning phenomenon in neural networks, which may be the underlying mechanism for why the loss spike improves generalization [18, 22], rather than simply controlling the value of $\lambda_{\max}$.

We believe that this work makes contributions in the following aspects:

(1) Analyzing the loss spike phenomenon and its frequency mechanism.

(2) Proposing the LLAS structure to explain the ascent stage of loss spikes.