An Experimental Study of Training Stability and Weight Sensitivity in Neural Networks

Siddharth Sankar Kakoti
Dhemaji Engineering College, Assam, India

Abstract

Understanding the stability and robustness of neural network training remains an important challenge in modern machine learning. While deep models often achieve high performance, their sensitivity to small perturbations in parameter space raises concerns regarding generalization and reliability. In this work, we conduct a controlled experimental study to analyze training stability, weight sensitivity, and local loss landscape behavior in neural networks. Using simple feedforward architectures trained on benchmark datasets, we examine how small perturbations in learned parameters affect model performance and how loss surface geometry correlates with observed robustness. Our results provide empirical evidence that models converging to sharper minima exhibit higher sensitivity to parameter perturbations, while flatter regions of the loss landscape are associated with improved stability.

1 Introduction

Neural networks have demonstrated remarkable success across a wide range of applications. Despite this success, the training dynamics and robustness properties of these models are not yet fully understood. In particular, it has been observed that models with similar training performance may differ significantly in their sensitivity to parameter perturbations and their ability to generalize.

Previous empirical studies suggest a relationship between the geometry of the loss landscape and model robustness, often described in terms of sharp versus flat minima. However, many of these observations remain qualitative, and further controlled experimental analysis is required to better understand these phenomena.

The objective of this study is to experimentally analyze:

- The sensitivity of trained neural networks to small perturbations in weight space
- The relationship between training stability and local loss landscape geometry
- How loss surface curvature relates to robustness under perturbations

2 Experimental Setup

2.1 Model Architecture

We use a fully connected feedforward neural network consisting of an input layer corresponding to flattened image inputs, two hidden layers with ReLU activations, and an output layer producing class logits. This architecture allows controlled experimentation while remaining representative of commonly used neural models.

2.2 Dataset

Experiments were conducted using the MNIST handwritten digit dataset. Input images were normalized using standard mean and variance normalization. The dataset was divided into training and test sets following standard benchmarks.

2.3 Training Procedure

Models were trained using the Adam optimizer with a fixed learning rate. Cross-entropy loss was used as the objective function. Training was performed for a fixed number of epochs to ensure convergence under standard conditions.

3 Weight Perturbation Analysis

To study weight sensitivity, we introduce additive Gaussian noise to trained model parameters after convergence. For a trained parameter vector W, perturbed parameters are defined as:

$$W' = W + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Multiple perturbation magnitudes were tested to observe how performance degrades as noise increases. Model performance was evaluated using classification accuracy and loss on the test dataset.

4 Loss Surface Visualization

To analyze local loss landscape geometry, we visualize the loss surface in randomly sampled two-dimensional subspaces of the parameter space. Two random direction vectors were sampled and normalized, and the loss was evaluated on a grid around the converged solution.

This approach provides insight into local curvature and the presence of sharp or flat regions surrounding the trained model parameters.

5 Results and Observations

5.1 Training Behavior

Under standard training conditions, the model exhibited stable convergence with smooth loss reduction. This baseline behavior served as a reference for subsequent perturbation experiments.

5.2 Sensitivity to Weight Perturbations

The following observations were made:

- Small perturbations had limited impact on performance
- Beyond a threshold perturbation magnitude, performance degraded rapidly
- Accuracy degradation was non-linear with respect to perturbation magnitude

5.3 Loss Landscape Geometry

Loss surface visualizations revealed notable differences in curvature around the converged solution. Regions exhibiting sharper curvature corresponded to higher sensitivity observed in perturbation experiments, while flatter regions were associated with greater robustness.

6 Discussion

The experimental results support the hypothesis that loss landscape geometry plays a significant role in determining model robustness. Sharp minima appear more sensitive to small perturbations, potentially impacting generalization and reliability.

While the models studied here are relatively simple, the observed trends align with broader empirical findings in deep learning research and motivate further investigation.

7 Limitations and Future Work

This study is limited to small-scale architectures and a single benchmark dataset. Future work includes:

- Extending analysis to deeper and more complex architectures
- Studying the effect of different optimization strategies
- Incorporating theoretical analysis to complement empirical findings
- Exploring connections between interpretability methods and training stability

8 Conclusion

In this work, we presented an experimental analysis of neural network training stability, weight sensitivity, and loss landscape behavior. Through controlled perturbation experiments and loss surface visualization, we demonstrated a clear relationship between local loss geometry and model robustness. These findings contribute to a better empirical understanding of training dynamics and highlight directions for further research.