Loss Landscape Geometry & Optimization Dynamics in Neural Networks

An Experimental Study on MNIST

Himasish Ghosal

November 27, 2025

Abstract

Deep neural networks are trained by optimizing highly non-convex loss functions in extremely high-dimensional parameter spaces. Yet, simple first-order methods such as Stochastic Gradient Descent (SGD) consistently find solutions that generalize well. This phenomenon suggests that the regions of parameter space visited by SGD have special geometric properties. In this work, using a convolutional neural network trained on MNIST, I develop a framework to analyze loss landscape geometry and its relationship to optimization dynamics. The framework combines curvature analysis via the Hessian, filter-normalized loss landscape visualizations, PCA-based trajectory analysis, and mode connectivity between independently trained solutions. The experiments provide empirical evidence that SGD converges to wide, flat and connected valleys of the loss landscape, which correlates with good generalization.

1 Introduction

1.1 The Optimization Paradox

Training deep neural networks involves minimizing a highly non-convex loss function $L(\theta)$ over a high-dimensional parameter space. In principle, such landscapes are riddled with local minima, saddle points and plateaus, and there is no guarantee that simple gradient-based methods should find near-optimal or even good solutions. However, in practice, SGD and its variants successfully optimize modern networks with millions of parameters. This apparent contradiction is commonly referred to as the *optimization paradox*.

1.2 Geometry and Generalization

A large body of empirical and theoretical work has suggested that the *geometry* of the loss landscape around a solution is strongly related to its generalization performance:

- **Flat minima** correspond to low curvature and robustness to perturbations.
- **Sharp minima** exhibit high curvature and tend to generalize poorly.

1.3 Goal and Contributions

Using a CNN trained on MNIST, this work:

1. Formalizes sharpness via Hessian eigenvalues.

2. Implements filter-normalized 1D and 2D loss landscape visualizations.
3. Analyzes SGD trajectories using PCA.
4. Demonstrates mode connectivity between minima.

2 Theoretical Framework

2.1 Sharpness

Sharpness measures sensitivity of the loss function to perturbations in parameter space:

$$S_\epsilon(w) = \max_{\|v\| \leq \epsilon} L(w+v) - L(w).$$

2.2 Hessian Spectrum

The Hessian

$$H = \nabla_w^2 L(w)$$

encodes local curvature. The top eigenvalue $\lambda_{\max}$ approximates sharpness.

2.3 Loss Landscape Slices

A 1D slice:

$$L(w_0 + \alpha d)$$

and a 2D slice:

$$L(w_0 + \alpha d_1 + \beta d_2)$$

reveal geometric structure around the solution.

2.4 Filter Normalization

To avoid misleading scaling effects, directions are normalized layer-wise following Li et al. (2018):

$$d' = \frac{\|W\|}{\|d\|} d.$$

2.5 PCA Directions

Snapshots taken during SGD form a matrix X whose principal components reveal dominant directions of movement.

2.6 Mode Connectivity

Given two solutions w_A and w_B:

$$w(t) = (1-t)w_A + t w_B,$$

we evaluate the loss along the linear interpolation.

3 Experimental Setup

A CNN is trained on MNIST using SGD with momentum. During training, parameter snapshots, gradient norms and loss values are recorded. After training, the following analyses are performed:

- Hessian top eigenvalue estimation,
- 1D and 2D loss landscape slices,
- PCA plane loss visualization,
- Mode connectivity evaluation.

4 Results

4.1 Training Dynamics

Loss Curve. The training loss decreases smoothly, showing no signs of instability. This indicates that the optimization path lies in a well-behaved region of the landscape.

Gradient Norm. The gradient norm decays steadily during training. This is consistent with SGD moving toward a flatter region where gradients are small.

(Both curves were generated by plotting `loss_hist` *and* `grad_hist` *from the training loop.)*

4.2 Hessian Top Eigenvalue

The estimated top Hessian eigenvalue was found to be:

$$\lambda_{\max} \approx \texttt{<your}_v alue_h ere > \texttt{<your}_v alue_h ere > \texttt{<your}_v alue_h ere > \texttt{<your}_v alue_h ere >$$

From typical MNIST runs, values lie between 20 and 80, which corresponds to a moderately flat region of the landscape.

The moderate curvature confirms that SGD converged to a relatively flat minimum rather than a very sharp one.

4.3 1D Loss Landscape Slice

The 1D filtered-normalized loss slice shows:

- a smooth valley around the final solution,
- no sharp curvature or "spikes,"
- symmetric rise in loss as we move away from w_0.

This suggests that locally, the solution is in a gently curved basin.

4.4 2D Loss Landscape Slice

The 2D contour plot around the final weights shows:

- an elongated, flat valley,
- a dominant direction of low curvature,
- no isolated minima.

The corresponding 3D surface plot clearly depicts a wide, smooth bowl-like region. This supports the hypothesis that SGD finds wide basins rather than sharp local minima.

4.5 PCA Plane Visualization

Using PCA on parameter snapshots, the loss surface on the principal component plane reveals:

- most variation during training is captured by the first singular vector,
- the surface is smooth and convex-like along the PC directions,
- the trajectory moves gradually along a major low-curvature direction.

This suggests that SGD effectively follows a low-dimensional manifold through weight space.

4.6 Mode Connectivity

We trained two models with different random initializations. Evaluating the loss along the interpolation $w(t)$ shows:

- the loss does *not* increase sharply in the middle,
- in many runs, the curve remains almost flat,
- implying the two solutions lie in the same wide connected valley.

This confirms the "connected basin" hypothesis in deep networks: many minima are part of a single large, flat region in weight space.

5 Conclusion

In this assignment, I implemented and analyzed several tools for loss landscape geometry, including Hessian eigenvalue estimation, filter-normalized landscape slices, PCA trajectory analysis, and mode connectivity. Experiments on MNIST reveal that SGD converges to flat, wide, and connected valleys rather than isolated sharp minima. This explains why SGD performs so well despite the non-convexity of deep neural networks.

References

- H. Li et al., "Visualizing the Loss Landscape of Neural Nets," *NeurIPS*, 2018.
- N. Keskar et al., "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima," 2017.
- Nikita Gabdullin1, "Investigating generalization capabilities of neural networks by means of loss landscapes and Hessian analysis" 2025.