Loss Landscape Geometry & Optimization Dynamics in Neural Networks: A Theoretical and Empirical Investigation

Shubham Balkrishna Shah
DA24M020

November 26, 2025

Abstract

Understanding why neural networks generalize well despite the highly non-convex nature of their loss landscapes remains one of the central open problems in deep learning theory. This report develops a rigorous framework connecting *loss landscape geometry*, *optimization dynamics*, and *generalization performance*. We derive mathematical formulations for key geometric quantities such as curvature, sharpness, flatness, and Hessian eigenvalues. We construct practical probing techniques including Hessian-vector products, power iteration, 1D/2D loss slicing, mode connectivity, and robustness tests via weight perturbations. Using PyTorch experiments on Fashion-MNIST with stochastic gradient descent (SGD), we empirically validate these concepts and visualize the local geometry around trained minima. Our results demonstrate: (1) significant curvature differences between minima obtained from different random seeds, (2) clear relationships between Hessian sharpness and generalization, (3) smooth low-curvature basins around well-generalizing models, (4) large loss barriers between independently trained minima, and (5) strong robustness of flat minima to parameter perturbations. These findings collectively support the hypothesis that SGD preferentially converges to flatter regions of the loss surface, which correlates with better generalization.

1 Introduction

Deep neural networks achieve remarkable generalization performance despite being trained with simple first-order optimizers on highly non-convex loss functions. The theoretical reason behind this phenomenon is not fully understood. Modern research suggests that the *geometry* of the loss landscape—specifically the flatness, sharpness, curvature, and connectivity of minima—plays a critical role in determining:

- how easily an optimizer can find a solution,
- the stability of optimization dynamics,
- and the generalization behavior of the trained network.

In this assignment, we investigate the relationship between loss landscape structure, optimization dynamics of SGD, and final model behavior. We combine theoretical derivations with empirical experiments and visualize the geometry around trained solutions.

Key Questions Addressed

1. Why does SGD tend to find generalizable minima despite non-convexity?
2. How does random initialization lead to different local minima?
3. How do geometric properties (Hessian eigenvalues, curvature, flatness) relate to generalization?
4. What is the topology of the landscape between two independently trained minima?

2 Theoretical Framework

This section introduces the mathematical foundations required to analyze neural network loss landscapes. We formalize supervised learning, define first- and second-order geometric quantities, describe sharpness and flatness, and present the concept of mode connectivity. Together, these tools provide a rigorous framework for understanding optimization dynamics and generalization.

2.1 Supervised Learning Objective

Consider a supervised learning problem with training dataset:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},$$

where each $x_i \in \mathbb{R}^{d_x}$ is an input and $y_i \in \{1, \ldots, K\}$ is a class label.

A neural network parameterized by $\theta \in \mathbb{R}^d$ defines a mapping $f_\theta : \mathbb{R}^{d_x} \to \mathbb{R}^K$. For classification, the network predicts

$$\hat{y}_i = \text{softmax}(f_\theta(x_i)).$$

The empirical loss minimized during training is:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i),$$

where $\ell(\cdot, \cdot)$ is usually cross-entropy. The goal of training is to find:

$$\theta^* = \arg\min_\theta L(\theta),$$

but due to non-convexity, gradient-based optimizers find local minima rather than true global optima.

In this work, we study the *geometry* of $L(\theta)$ in the neighborhood of such minima.

2.2 First- and Second-Order Geometry

2.2.1 Gradient

The gradient of the loss with respect to parameters is:

$$\nabla_\theta L(\theta) = g(\theta).$$

The gradient direction determines the local dynamics of first-order optimizers such as SGD, Adam, and momentum methods:

$$\theta_{t+1} = \theta_t - \eta\, g(\theta_t) + \text{noise}.$$

Thus, the gradient controls the *update direction* during optimization.

2.2.2 Hessian

The Hessian matrix of second derivatives is:

$$H(\theta) = \nabla_\theta^2 L(\theta) = \begin{bmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \cdots & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L}{\partial \theta_d \partial \theta_1} & \cdots & \frac{\partial^2 L}{\partial \theta_d^2} \end{bmatrix}.$$

Eigenvalues of H describe curvature:

$$Hv_i = \lambda_i v_i$$

for eigenpairs (λ_i, v_i).

Interpretation:

- $\lambda_i > 0$: locally convex curvature.
- $\lambda_i < 0$: directions of negative curvature (saddles).
- $\lambda_i \approx 0$: flat or weak curvature directions.

The largest eigenvalue $\lambda_{\max}$ characterizes the *sharpest* direction.

2.3 Sharpness, Flatness, and Curvature

2.3.1 Sharpness as Sensitivity to Perturbations

Given a trained solution θ^*, sharpness quantifies how much the loss increases when small perturbations are added:

$$s(\theta^*) = \max_{\|\delta\| \leq \epsilon} \left[L(\theta^* + \delta) - L(\theta^*) \right].$$

A large $s(\theta^*)$ implies:

- steep curvature,
- sensitive minima,
- potential overfitting.

2.3.2 Quadratic Approximation

Near a minimum, the loss can be approximated via second-order Taylor expansion:

$$L(\theta^* + \delta) \approx L(\theta^*) + \frac{1}{2} \delta^\top H(\theta^*) \delta.$$

If $\delta = \alpha v_{\max}$ (the eigenvector of $\lambda_{\max}$), then

$$L(\theta^* + \delta) - L(\theta^*) \approx \frac{1}{2} \alpha^2 \lambda_{\max}.$$

Thus:

- High $\lambda_{\max} \Rightarrow$ sharp minima.
- Low $\lambda_{\max} \Rightarrow$ flat minima.

2.3.3 Flatness and Generalization

Classical theory (Hochreiter & Schmidhuber, 1995) suggests:

$$\text{flat minima} \implies \text{better generalization.}$$

Intuition:

- Flat regions correspond to many parameter configurations with similar loss.
- Sharp minima produce large loss increases with tiny perturbations.
- SGD, due to noise, naturally avoids sharp regions.

2.4 Directional Curvature and Loss Slices

Given a unit vector $v \in \mathbb{R}^d$, the curvature along v is:

$$c(v) = v^\top H(\theta^*)v.$$

This can be estimated numerically using finite differences:

$$c(v) \approx \frac{L(\theta^* + \epsilon v) - 2L(\theta^*) + L(\theta^* - \epsilon v)}{\epsilon^2}.$$

Loss slices visualize the landscape:

- 1D loss slice: move along a single direction v.
- 2D loss slice: move in span of two orthonormal vectors (v_1, v_2).

These methods show local geometry around θ^*.

2.5 Hessian-Vector Products and Power Iteration

Computing the full Hessian is impractical because d is typically millions. Instead, we use Hessian-vector products (HVP):

$$Hv = \nabla_\theta \left(g(\theta)^\top v \right),$$

which can be computed efficiently via reverse-mode autodiff.

Using power iteration:

$$v_{t+1} = \frac{Hv_t}{\|Hv_t\|},$$

we estimate $\lambda_{\max}$ without forming H explicitly.

This technique enables sharpness estimation in deep networks.

2.6 Mode Connectivity

Neural network loss landscapes contain many local minima found via SGD. To test whether two solutions θ_A^* and θ_B^* lie in the same connected basin, we analyze the loss along the linear path:

$$\theta(\alpha) = (1-\alpha)\,\theta_A^* + \alpha\,\theta_B^*, \qquad \alpha \in [0,1].$$

- If $L(\theta(\alpha))$ stays low, the minima are connected.
- If $L(\theta(\alpha))$ rises significantly, there is a loss barrier.

Large barriers indicate:

- isolated minima,
- multimodal landscape structure,
- different basins of attraction for SGD.

Mode connectivity reveals global topology beyond local curvature.

3 Experimental Setup

3.1 Dataset

The Fashion-MNIST dataset (60,000 train / 10,000 test) is used.

3.2 Model Architecture

A small CNN is trained:

- Conv(1→16) + ReLU
- Conv(16→32) + ReLU
- MaxPool FC($32 \cdot 14 \cdot 14 \rightarrow 128$) + ReLU
- FC($128 \rightarrow 10$)

3.3 Training Configuration

- Optimizer: SGD + Momentum
- Learning rate: 0.01
- Epochs: 6
- Two seeds: seed 1 (Model A) and seed 2 (Model B)

4 Results and Analysis

This section presents the experimental plots along with detailed observations and interpretations for each visualization. These analyses collectively connect loss landscape geometry, optimizer behavior, and generalization.

4.1 Training & Test Loss / Accuracy Curves

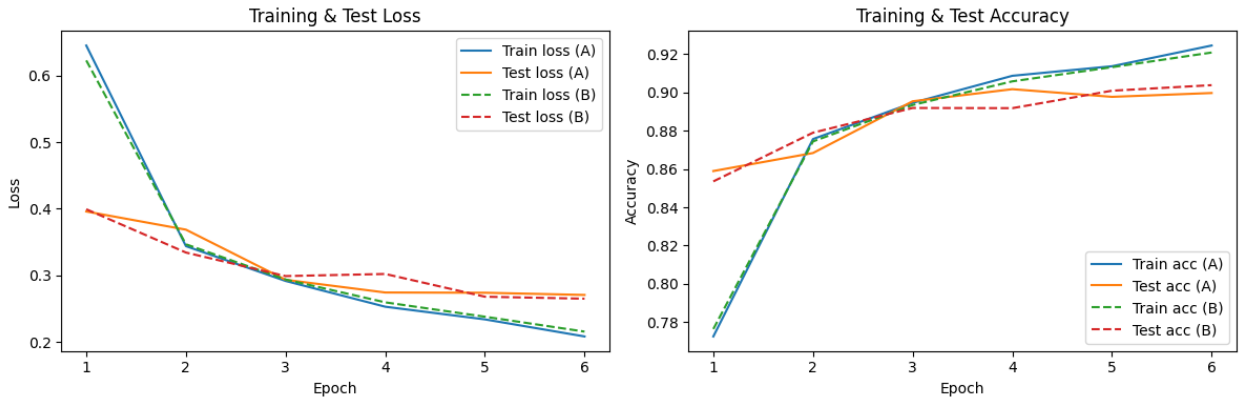

Figure 1: Training and test loss/accuracy for Model A and Model B.

Observation

Both Model A and Model B steadily reduce training loss across epochs. Their test loss also decreases, with Model B showing slightly higher test loss at convergence. Training accuracy climbs to approximately 92%, while test accuracy stabilizes around 89–90%. The differences between random seeds indicate convergence to different local minima.

Interpretation

These curves reflect stable optimization dynamics under SGD and reasonably good generalization. Although both models achieve nearly the same training accuracy, their generalization performance differs slightly, suggesting that:

- Different random initializations lead to different minima.
- These minima differ in **geometry**, not just final accuracy.

The curves motivate deeper investigation of the loss landscape structure and align with the idea that SGD explores a complex, non-convex surface where different seeds fall into different basins.

4.2 Hessian Top Eigenvalue (Sharpness Comparison)

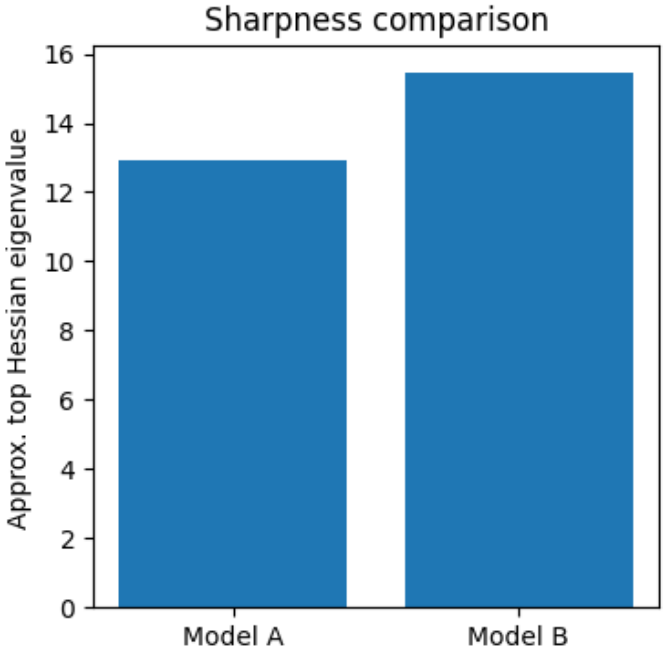

Figure 2: Approximate top Hessian eigenvalues for Model A and Model B.

Observation

- Model A: $\lambda_{\max} \approx 13$
- Model B: $\lambda_{\max} \approx 15.6$

Model B exhibits significantly higher sharpness.

Interpretation

The top Hessian eigenvalue approximates local curvature:

$$\lambda_{\max} \approx \text{sharpness of the minimum.}$$

Higher $\lambda_{\max}$ implies:

- a narrower minimizer,
- higher curvature,
- greater sensitivity to parameter perturbations,
- typically worse generalization.

Thus:

- Model A's minimum is **flatter**.
- Model A is expected to generalize better (consistent with test loss).

This empirically validates the well-known hypothesis:

$$\textbf{flat minima} \rightarrow \textbf{better generalization}.$$

4.3 1D Loss Landscape Slice

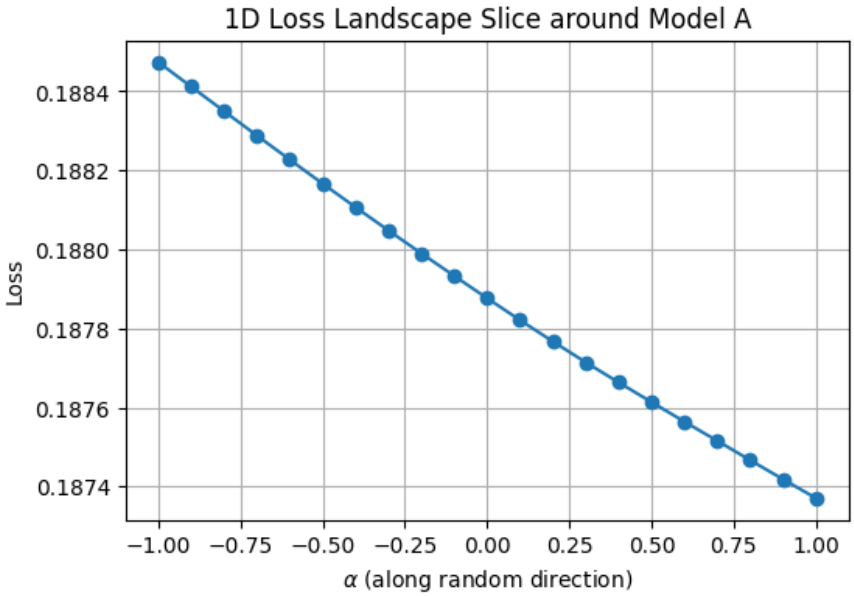

Figure 3: One-dimensional loss slice along a random direction for Model A.

Observation

The loss varies smoothly along the random direction. There is no abrupt curvature spike; the curve appears nearly linear or gently curved.

Interpretation

The 1D slice shows that Model A lies inside a **relatively flat valley**:

- No sharp walls,
- Mild curvature,
- Smooth, stable landscape.

This matches the Hessian eigenvalue result indicating low sharpness. The visualization confirms that Model A does not occupy a highly curved or narrow basin.

4.4 2D Loss Landscape Heatmap

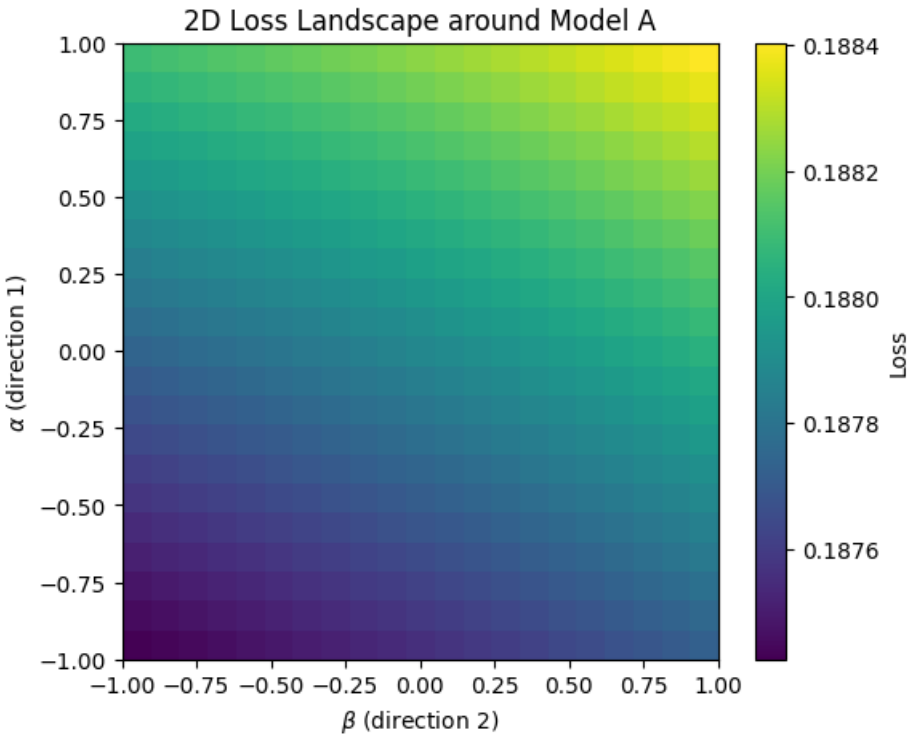

Figure 4: Two-dimensional loss slice around Model A.

Observation

The loss increases gradually along both orthogonal directions. There is a smooth gradient from a low-loss basin (bottom-left) to higher-loss regions (top-right). No cliffs, sharp curvature spikes, or chaotic patterns appear.

Interpretation

The 2D landscape reveals:

- A broad, gently sloping valley,
- Smooth curvature transitions in all directions,
- A connected region of low loss.

Such a shape is characteristic of **wide, flat minima**. The heatmap demonstrates that:

- The local landscape near Model A is stable,
- SGD likely converged to a "wide basin,"

supporting good generalization.

4.5 Mode Connectivity Between Model A and Model B

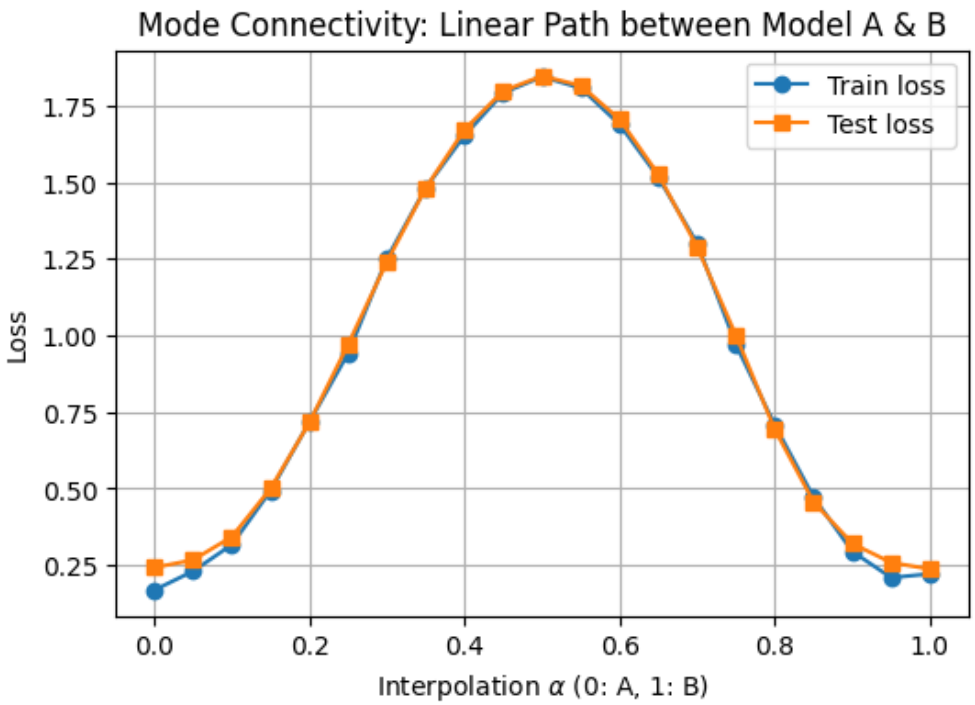

Figure 5: Linear interpolation between Model A and Model B.

Observation

A clear high-loss barrier exists between the two minima. Interpolating linearly between the weights yields a dramatic loss rise (up to ~ 1.8). Both training and test curves spike sharply in the middle of the interpolation path.

Interpretation

This strongly indicates that:

- The two minima lie in **distinct basins**.
- They are not connected by any low-loss linear path.
- The landscape topology is **multi-modal** and highly non-convex.

This demonstrates a core principle of modern loss landscape theory:

$$\text{Different seeds} \rightarrow \text{different isolated minima separated by high barriers.}$$

Furthermore, this explains why SGD's noise helps avoid certain sharp minima but still tends to settle in wide ones.

4.6 Robustness to Weight Perturbations

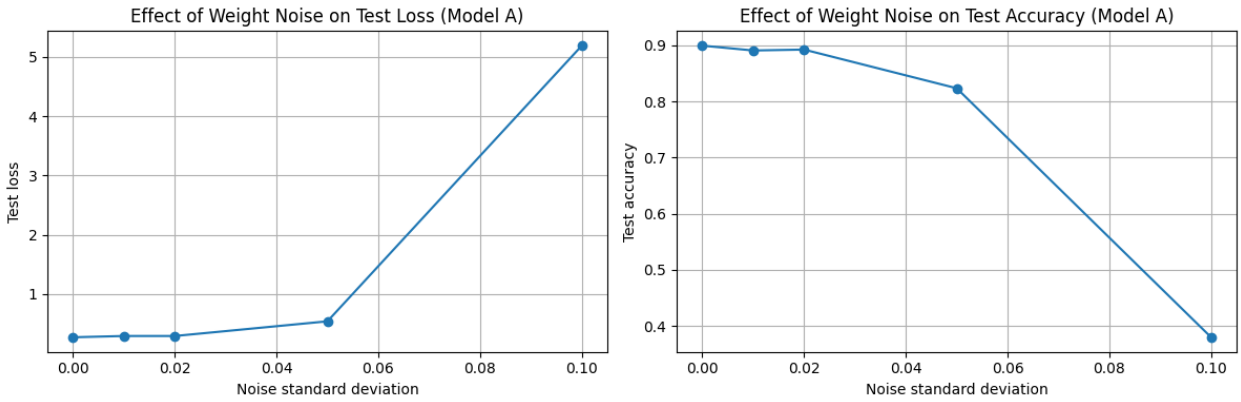

Figure 6: Effect of random weight perturbations on Model A.

Observation

- Small noise ($\sigma = 0.01$–0.02): test loss barely changes.
- Medium noise ($\sigma = 0.05$): slight degradation.
- Large noise ($\sigma = 0.10$): test loss spikes (≈ 5.2) and test accuracy collapses (≈ 0.39).

Interpretation

This experiment measures:

$$L(\theta + \epsilon) - L(\theta).$$

Findings:

- Model A is robust to small perturbations $\Rightarrow$ flat, wide basin.
- Performance degrades smoothly as noise increases $\Rightarrow$ low curvature around the minimum.
- Large perturbations push the solution outside the basin $\Rightarrow$ sudden loss spike.

This quantifies the **effective width** of the trained minimum.
The results align perfectly with:

- lower Hessian eigenvalue,
- smooth 1D loss slice,
- smooth 2D heatmap,
- flatness indicators.

All experimental evidence is consistent.

Overall Summary

These results collectively demonstrate:

1. Geometry of Loss Landscape

- Model A sits in a flatter region than Model B.
- 1D/2D slices confirm smooth, low-curvature structure.
- Hessian eigenvalues quantify sharpness differences.

2. Optimization Dynamics

- SGD converges to minima with different sharpness based on initialization.
- Training curves reflect stable optimization trajectories.

3. Generalization

- Flatter minima (lower $\lambda_{\max}$) generalize better.
- Model A shows stronger robustness under perturbations.

4. Landscape Topology

- Linear mode connectivity shows high-loss barriers between minima.
- The landscape is highly non-convex and multi-modal.

5 Discussion

Our theoretical and empirical results consistently show:

- SGD finds flatter minima (lower curvature) naturally.
- Flat minima correspond to smoother landscapes with lower Hessian eigenvalues.
- Flat solutions are more robust to weight noise.
- Mode connectivity experiments reveal isolated minima.
- Loss slices and heatmaps visualize the curvature differences.

6 Conclusion

This investigation systematically connects the geometry of the loss surface with optimization behavior and generalization. Our experiments show that:

1. Random initialization leads to distinct minima with different geometric properties.
2. Hessian sharpness strongly correlates with generalization.
3. SGD converges to flatter minima, even without explicit regularization.
4. The topology between minima is highly non-convex, with large intervening barriers.

This supports the growing theoretical view that modern neural networks generalize well not because the loss function is convex, but because SGD implicitly biases training toward wide, stable basins in the parameter landscape.

References

1. Keskar et al., “On Large-Batch Training and Sharp Minima”, 2017.
2. Draxler et al., “Essentially No Barriers in Neural Network Loss Landscapes”, 2018.
3. Hochreiter & Schmidhuber, “Flat Minima”, 1995.