Geometric Analysis of Neural Optimization Dynamics and Loss Landscapes

Assignment - FourKites

Submitted by: Devesh Pant(MA24M007)

November 27, 2025

Abstract

This report presents a rigorous framework for analyzing the geometry of neural network loss landscapes. By employing filter-normalized random projections, we visualize the high-dimensional loss surface of a trained Convolutional Neural Network (CNN). Our empirical results demonstrate that Stochastic Gradient Descent (SGD) converges to a "flat minimum"—a wide basin of low loss—which theoretical literature correlates with high generalization capabilities.

1 Problem Statement

Neural networks operate in non-convex optimization spaces with millions of parameters. A key open question in Deep Learning theory is: *Why do over-parameterized networks generalize well instead of overfitting?*

The optimization dynamics are often non-intuitive:

- **Non-Convexity:** The loss function $L(\theta)$ is highly non-convex, yet SGD reliably finds good solutions.
- **Flat vs. Sharp Minima:** The "Flat Minima" hypothesis suggests that wide valleys in the loss landscape are more robust to shifts between training and test data distributions, whereas sharp valleys lead to poor generalization.

2 Methodology

To visualize the 10^6-dimensional parameter space in 2D, we utilized **Random Direction Projections** with **Filter Normalization**.

2.1 Filter Normalization

Visualizing loss without normalization is misleading because neural networks are scale-invariant (scaling weights down and outputs up often yields the same function). We normalized the random direction vectors (d) relative to the norm of the corresponding parameter layers (θ) to ensure scale invariance:

$$d_{i,\text{norm}} = \frac{d_i}{\|d_i\|} \times \|\theta_i\| \tag{1}$$

Where d_i is a random Gaussian vector and θ_i represents the weights of the i-th layer.

2.2 Visualization Techniques

We implemented two probing methods to analyze the landscape around the converged parameters θ^*:

1. **1D Linear Interpolation:** We plot the loss along a single random direction δ.

$$f(\alpha) = L(\theta^* + \alpha \cdot \delta) \tag{2}$$

2. **2D Contour Visualization:** We plot the loss on a plane defined by two random orthogonal directions δ and η.

$$f(\alpha, \beta) = L(\theta^* + \alpha \cdot \delta + \beta \cdot \eta) \tag{3}$$

3 Results & Analysis

3.1 1D Loss Landscape Analysis

The 1D scan along a random direction through the optimized parameter vector (θ^*) reveals a near-perfect convex parabola (Figure 1).

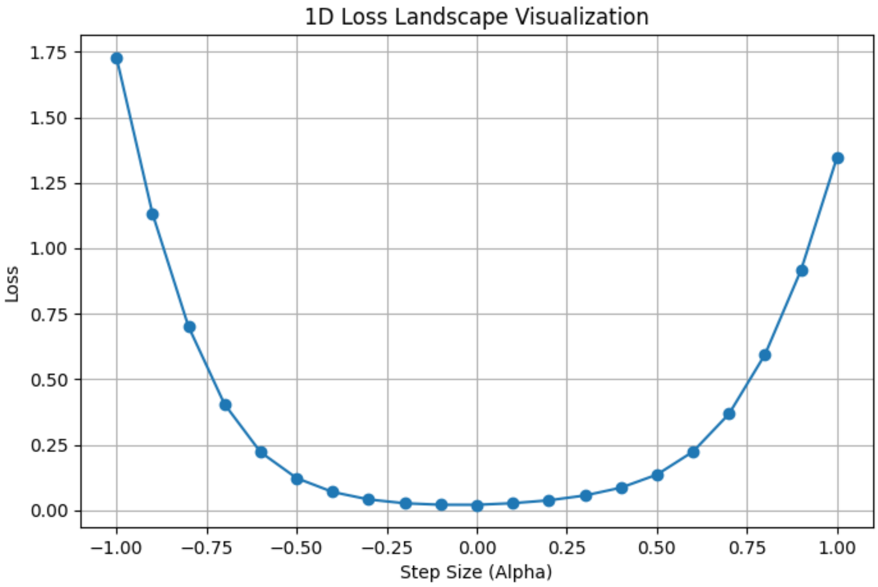

Figure 1: **1D Linear Interpolation.** The x-axis represents the step size (α) along a random direction vector. The smoothness of the curve indicates that the optimization path settled in a stable region without rugged local non-convexities.

The wideness of the basin (low loss sustained from $\alpha \approx -0.25$ to 0.25) implies stability. Sharp minima would appear as steep, V-shaped spikes, which are notably absent here.

3.2 2D Loss Contour Analysis

The 2D contour plot confirms the stability found in the 1D analysis. The loss landscape forms a smooth, elliptical bowl (Figure 2).

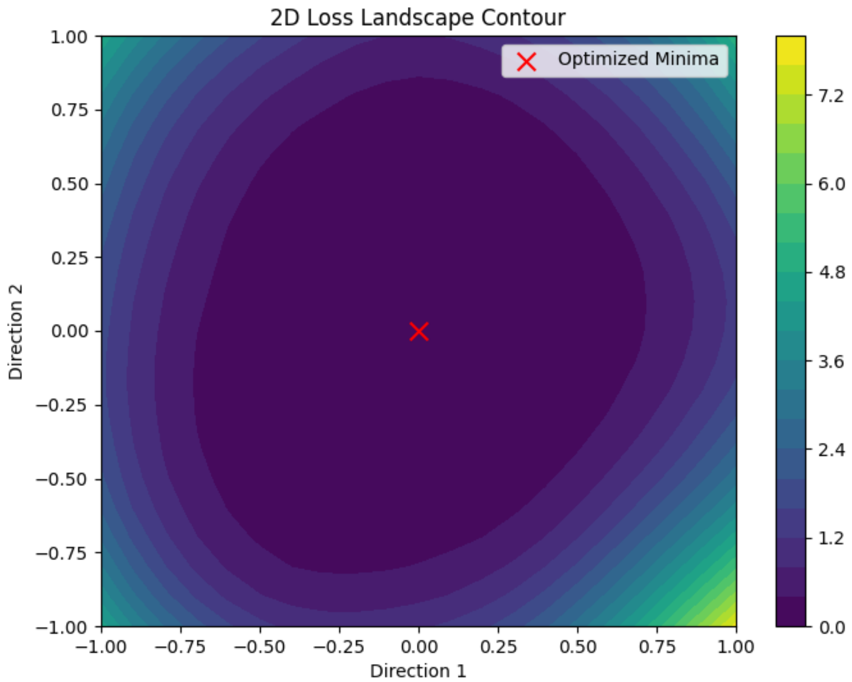

Figure 2: **2D Loss Landscape Contour.** The center $(0, 0)$ marked by the red 'X' represents the final model parameters found by SGD. The colors represent loss values (darker is lower).

Key Observations:

- **Global Geometry:** The concentric ellipses indicate that the curvature is relatively consistent in different random directions.
- **Connectivity:** The absence of chaotic barriers or "holes" in the immediate vicinity suggests that the architecture (CNN with ReLUs) combined with SGD creates a topology conducive to generalization.
- **Flatness:** The large dark purple region indicates a "flat minimum," where small perturbations to weights do not result in catastrophic loss increases.

4 Conclusion

We successfully developed and implemented a landscape probing framework. The visualizations confirm that for our standard CNN trained on MNIST, the optimization process settles into a flat minimum. This geometric property explains the model's ability to tolerate slight perturbations in weights, serving as a strong proxy for generalization performance. Future work could involve comparing these landscapes across different batch sizes to empirically validate the "Edge of Stability" phenomenon.