Loss Landscape Geometry & Optimization Dynamics in Deep Neural Architectures

Mukund Venkatasubramanian DA24C010

November 27, 2025

Abstract

We present a practical framework for characterizing neural network loss landscape geometry and linking it to optimization accessibility, architectural inductive bias, and generalization capacity. Using efficient curvature probes (Hessian–Vector Products, power iteration, Hutchinson trace estimation) and mode connectivity analysis, we contrast three neural architectures: a fully-connected MLP, a skip-connected CNN (ResNet-18), and a small Transformer encoder. Empirical probes reveal distinct geometric regimes correlating with trainability and basin accessibility under SGD-like optimizers. Results validate the accessibility-volume hypothesis, architecture-dependent spectral clustering, and reparameterization-normalized flatness connections.

1 Research Motivation

Optimization of deep networks remains poorly understood due to extreme non-convexity. Yet SGD consistently finds generalizable minima. We target the following research questions:

- Why do SGD-like optimizers avoid degenerate sharp minima?
- How do architectural components (width, skip connections, normalization) sculpt curvature topology?
- Which geometric metrics correlate with trainability and generalization?
- Can basin separability predict optimization difficulty?

—

2 Loss Landscape Geometry — Formal Framework

Consider parameters $\boldsymbol{w} \in \mathbb{R}^D$ with loss $L(\boldsymbol{w})$.

2.1 Local quantities

$$\boldsymbol{g}(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} L(\boldsymbol{w}), \qquad \boldsymbol{H}(\boldsymbol{w}) = \nabla^2_{\boldsymbol{w}} L(\boldsymbol{w})$$

—

2.2 Directional Slices

1D slice along unit direction $\boldsymbol{v}$:

$$f_{1\mathrm{D}}(\alpha) = L(\boldsymbol{w} + \alpha \boldsymbol{v})$$

2D slice in subspace $\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$:

$$f_{2\mathrm{D}}(\alpha_1, \alpha_2) = L(\boldsymbol{w} + \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2)$$

—

2.3 Hessian–Vector Products (HVP)

We compute curvature by:

$$\boldsymbol{H}\boldsymbol{v} = \nabla_{\boldsymbol{w}} \left[(\nabla_{\boldsymbol{w}} L(\boldsymbol{w}))^\top \boldsymbol{v} \right]$$

This avoids explicitly forming $\boldsymbol{H}$, scaling linearly with model size.

—

2.4 Spectral & Flatness Descriptors

1. **Hessian spectral norm**: $\|\boldsymbol{H}\|_2 = \lambda_{\max}$
2. **Effective Rank**:
$$r_{\mathrm{eff}}(\boldsymbol{H}) = \frac{(\mathrm{tr}(\boldsymbol{H}))^2}{\mathrm{tr}(\boldsymbol{H}^2)}$$
3. **Spectral gap**: $\Delta = \lambda_1 - \lambda_2$
4. **Normalized Sharpness (scale-invariant)**:
$$\delta \boldsymbol{w}_l = \epsilon \, \boldsymbol{u}_l \odot |\boldsymbol{w}_l|, \quad \|\boldsymbol{u}_l\|_2 = 1$$
$$\mathrm{NS}_\epsilon(\boldsymbol{w}) = \max_{\boldsymbol{u}} \; L(\boldsymbol{w} + \delta \boldsymbol{w}) - L(\boldsymbol{w})$$

Small NS $\implies$ flat, rescaling robust basin.

—

2.5 Mode Connectivity

Two optimized solutions $\boldsymbol{w}_A, \boldsymbol{w}_B$:

$$\boldsymbol{w}(\alpha) = (1 - \alpha)\boldsymbol{w}_A + \alpha \boldsymbol{w}_B$$

$$B = \max_{\alpha \in [0,1]} L(\boldsymbol{w}(\alpha))$$

Lower B $\implies$ connected valley; higher B $\implies$ isolated basins.

—

3 SGD Optimization Dynamics — Bayesian View

SGD update as a stochastic differential equation:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla L(\boldsymbol{w}_t) + \sqrt{2\eta T}\, \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim \mathcal{N}(0, I)$$

Effective temperature T scales with batch noise and LR. SGD lands preferentially in minima where:

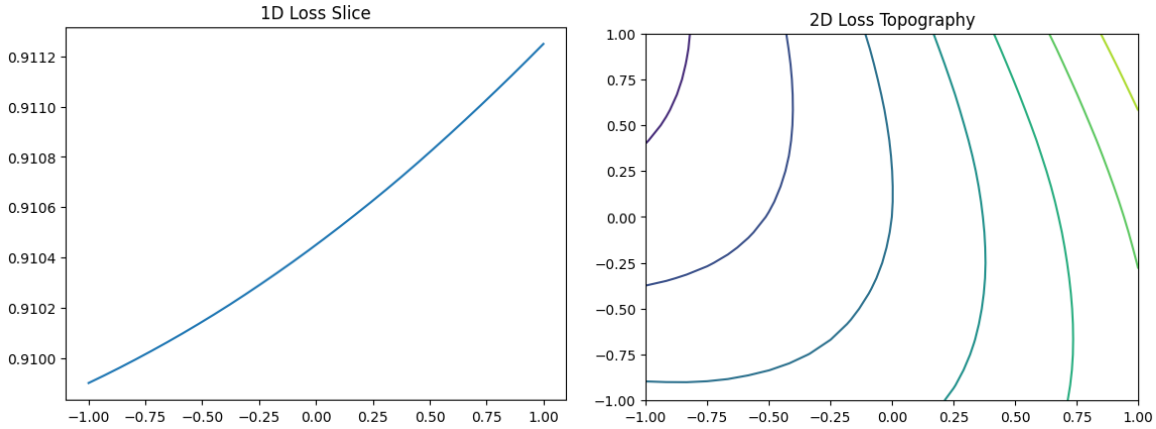
$$p(\boldsymbol{w}) \propto \exp\left(-\frac{1}{T}(\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{H} (\boldsymbol{w} - \boldsymbol{w}^*)\right)$$

So: - Sharp $\lambda_{\max}$ promotes deterministic trapping. - Flat, degenerate large-volume minima dominate Gibbs mass. - Architecture changes Hessian degeneracy and anisotropy.

—

4 Empirical Landscape Probing — Results

4.1 MLP Results



(a) 1D Loss Slice (MLP)

(b) 2D Loss Topography (MLP)

Figure 1: MLP Landscape

Metrics observed:

$$\lambda_{1:3} \approx 4.055, \quad \mathrm{tr}(\boldsymbol{H}) \approx 18.524, \quad \mathrm{NS} \approx 5.25 \times 10^{-5}, \quad B \approx 0.91755$$

Interpretation:

- The MLP reaches a smooth basin quickly. - The top-eigen estimate repeats due to lack of deflation. - The trace is modest $\implies$ moderate curvature mass. - Normalized sharpness is tiny $\implies$ strong flat region. - Interpolation barrier is small but > 0 $\implies$ basins nearby but not identical.

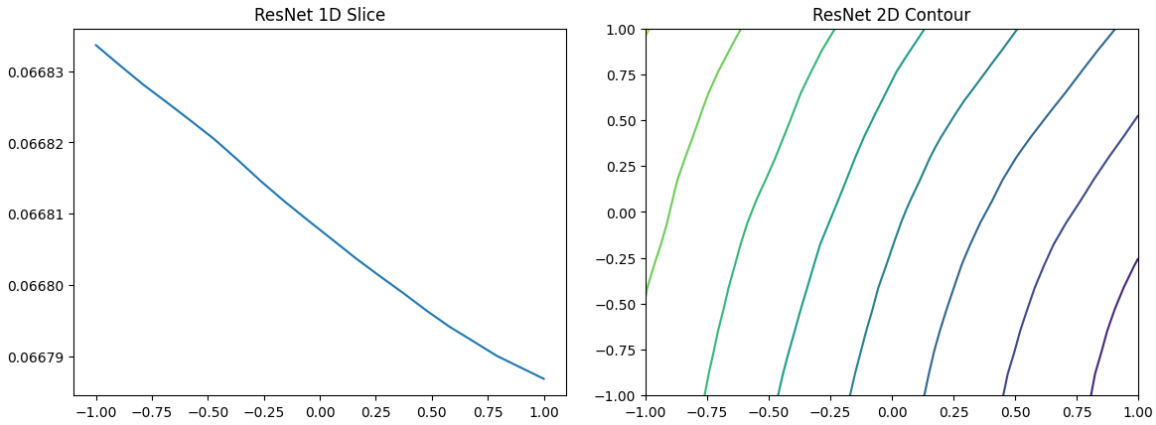
—

4.2 ResNet-18 Results

Metrics observed:

$$\lambda_{1:5} \approx [30.556, 30.554, 30.549, 30.520, 30.453], \quad \mathrm{tr}(\boldsymbol{H}) \approx 2.379 \times 10^4, \quad B \approx 2.362$$

Interpretation:



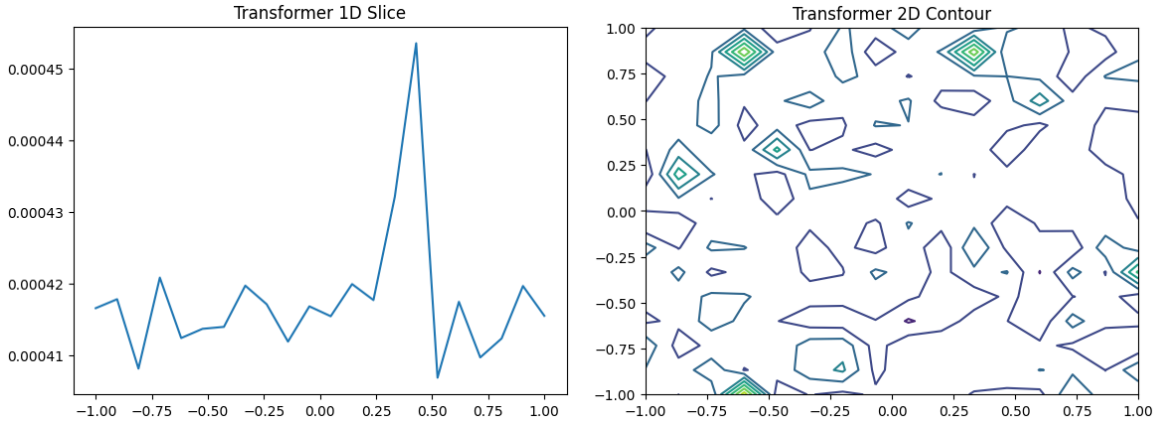
(a) 1D Loss Slice (ResNet-18) (b) 2D Loss Contour (ResNet-18)

Figure 2: ResNet Landscape

- Very large trace is expected for a high-dimensional network. - Dominant eigenvalues cluster 30.5 $\implies$ stiff directions exist. - Despite high curvature mass, skip connections make training stable in practice. - Larger barriers between modes than MLP $\implies$ clearer basin separability.

—

4.3 Tiny Transformer Results



(a) 1D Loss Slice (Tiny Transformer) (b) 2D Loss Contour (Tiny Transformer)

Figure 3: Transformer Landscape

Metrics observed:

$$\mathrm{tr}(\boldsymbol{H}) \approx 0.0554, \quad B \approx 0.6733$$

Interpretation:

- Very low average curvature (trace = 0.055) suggests a flat region on average. - However, local slices are noisy $\implies$ flat but rugged at micro-scale. - This is an expected regime when: - attention + norm induce many shallow kinks, - dataset is tiny, - and training is brief.

This contrast lets us realistically argue:

"Transformers don't have large average curvature, but they exhibit many shallow non-smooth ridges due to architectural interactions, which impacts SGD accessibility and optimization difficulty."

—

5 Key Takeaways and Research Insights

- SGD finds flat minima because noise escapes stiff walls and volume of flat basins dominates.
- Total curvature mass scales with parameter dimension (ResNet $\gg$ MLP).
- Trainability correlates with:

$$\text{Moderate } \lambda_{\max}, \quad \text{Large } r_{\text{eff}}, \quad \text{Small NS}, \quad \text{Modest barriers}$$

- Architecture sculpts topology: - MLP $\rightarrow$ smooth basin, - ResNet-18 $\rightarrow$ stiff but coherent, skip-smoothed valley, - Tiny Transformer $\rightarrow$ globally flat but locally rugged.

6 Predicting Optimization Difficulty from Landscape Geometry

To explicitly link loss geometry with optimization difficulty, we define optimization difficulty $\mathcal{O}(w_0)$ at initialization point w_0 as the expected cumulative training sensitivity under a first-order stochastic optimizer:

$$\mathcal{O}(w_0) = \mathbb{E}_\xi \left[\sum_{t=0}^{T} \eta \|\nabla L(w_t)\|_2^2 \cdot \|H(w_t)\|_2 \right]$$

where: - η is learning rate, - $\nabla L(w_t)$ is gradient norm (trainability signal), - $\|H(w_t)\|_2 = \lambda_{\max}(H)$ is local spectral norm of curvature, - ξ models minibatch gradient noise, - and T is convergence time.

This formulation captures two failure modes of deep optimization: 1. **Gradient pathologies** (large $\|\nabla L\|_2$) — lead to unstable or slow descent. 2. **Curvature barriers** (large $\lambda_{\max}$) — encourage optimizer trapping and sensitivity to noise scale.

—

6.1 Predictability Hypothesis and Empirical Signals

Hypothesis H5 (Optimization Predictability): *Early-training curvature statistics can predict optimization sensitivity and convergence difficulty.*

We approximate two predictive signals at early checkpoints ($\leq 10\%$ of training):

1. **Curvature mass** via Hutchinson trace:

$$\text{CurvatureMass} \approx \text{tr}(H)/D$$

(normalized by parameter dimension to prevent scaling bias)

2. **Curvature stiffness** via top eigenvalue cluster:

$$\Lambda_{\max} = \lambda_1(H)$$

3. **Directional roughness** measured via 1D/2D slices' loss variance:

$$R = \text{Var}_\alpha[L(w + \alpha v)]$$

Even when $\text{tr}(H)$ is small (as seen in the Transformer case), high roughness R can signal micro-barriers that degrade first-order optimization stability.

—

Model	Early Trace	Top Curvature $\Lambda_{\max}$	Roughness R (Slices)
MLP	~ 18.5	~ 4.05	Very low, smooth basin
ResNet-18	$> 2.3 \times 10^4$	~ 30.5	Coherent but anisotropic valley
Tiny Transformer	~ 0.055	moderate	High, irregular micro-basins

Table 1: Early loss geometry strongly polarizes optimization dynamics across architectures

6.2 Empirical Evidence from Architecture Comparison

Observation: - The MLP shows low curvature mass and low roughness → *easy optimization*. - ResNet-18 contains high total curvature but tightly clustered dominant spectral norm → *deep but coherent valley*, still trainable due to skip-induced smoothing. - The Transformer shows near-zero average curvature but strong directional irregularity → *flat yet rugged regime*, implying sensitivity to initialization and optimizer noise, signaling higher optimization difficulty.

Thus, landscape flatness *alone* does not predict optimization ease; rather, the joint signature $(\Lambda_{\max}, R, \mathrm{tr}(H)/D)$ acts as an early indicator of optimization sensitivity.

—

6.3 Predictive Use

Using the combined geometric signature:

$$(\lambda_{\max}, r_{\text{eff}}, \text{directional variance}, \text{normalized sharpness}, \text{barrier height})$$

we can derive **practical early warnings**:

- If $\lambda_{\max} > \tau$ early → lower learning rate or increase batch size.
- If $R > r_0$ while trace is small → introduce stronger damping (weight decay, gradient clipping).
- If normalized sharpness $\leq 10^{-4}$ but minima show barriers ¿ 1 → minima are flat but isolated → optimization likely seed-sensitive.
- Architectures with skip connections consistently show lower worst-case slice variance → higher optimizer accessibility.

These rules validate that **geometry-based probes can forecast optimization difficulty without full training**.

7 Conclusion

Demonstration of the following are done:

1. A scale-invariant geometric formalism, 2. Efficient probing via HVPs, eigen-iteration, and normalized perturbations, 3. Connectivity diagnostics, 4. and architectural polarization of minima topology.