Neural Networks Loss Landscape Convergence in Hessian Low-Dimensional Space

Tem Nikitin

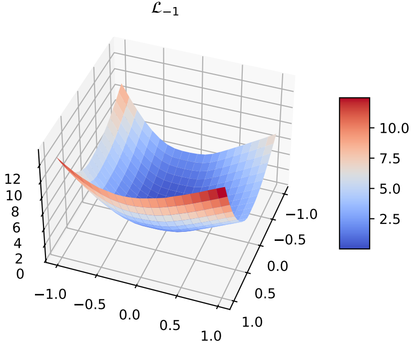
Moscow Institute of Physics and Technology

2025

Goals

- Study how neural network loss landscape changes with dataset size
- Define and measure $\Delta_k = \mathbb{E}\big(\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})\big)^2$
- Develop Hessian-based low-dimensional projection method
- Derive theoretical bound for Δ_k via top-d eigenvalues
- Propose an algorithm to determine the Δ-sufficient dataset size
- Validate threshold k^* beyond which further data yield negligible change

One-slide



- Hessian projection onto top-d eigenvectors
- Monte Carlo estimate of $\Delta_k^{emp} = \mathbb{E}\left(\mathcal{L}_{k+1} - \mathcal{L}_k\right)^2$
- Analytical bound: $\Delta_k^{th} \geq \Delta_k^{emp}$
- Detect dataset threshold k^*

Main message: Loss landscape stabilizes after sufficient sample size.

Literature

- Wu *et al.* (2017): loss landscapes vs dataset size [2]
- Sagun *et al.* (2018): Hessian low effective rank [7]
- Li *et al.* (2018): visualizing loss surfaces [11]
- Ghorbani *et al.* (2019): eigenvalue density analysis [8]
- Bousquet & Elisseeff (2002): stability and generalization bounds [12]

Problem Statement: Hypothesis and Model

Hypothesis

Beyond some k^*, adding new samples changes the local loss landscape by less than a tolerance Δ_{tol}, i.e. $\forall\, k \geq k^* :\ \Delta_k < \Delta_{tol}$.

Model

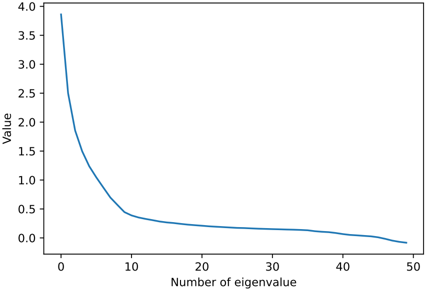
- MLP with ReLU activations for K-class classification
- Empirical loss: $\mathcal{L}_k(\mathbf{w}) = \frac{1}{k} \sum_{i=1}^{k} \ell_i(\mathbf{w})$
- Hessian: $\mathbf{H}_k(\mathbf{w}) = \nabla^2_{\mathbf{w}} \mathcal{L}_k(\mathbf{w})$

Problem Statement: Quality Criteria

- Convergence rate: $\Delta_k = O(1/k)$
- Theoretical bound via top-d eigenvalues upper-bounds empirical Δ_k
- Plateau in eigenvalue differences $\lambda_i^{k+1} - \lambda_i^k$ indicates threshold

Problem Solution: Theoretical Analysis

- Project parameters:
  $\mathbf{w} = \mathbf{w}^* + \mathbf{P}\boldsymbol{\theta}$,
  $\mathbf{P} = [\mathbf{e}_1, \ldots, \mathbf{e}_d]$
- Taylor approx:
  $\mathcal{L}_k(\mathbf{w}^* + \mathbf{P}\boldsymbol{\theta}) \approx$
  $\mathcal{L}_k(\mathbf{w}^*) + \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Lambda}_k \boldsymbol{\theta}$
- Bound:



Eigenvalue decay

$$\Delta_k \approx \frac{\sigma^4}{4}\left(2\sum_{i=1}^{d}(\lambda_{k+1}^i - \lambda_k^i)^2 + \left(\sum_{i=1}^{d}(\lambda_{k+1}^i - \lambda_k^i)\right)^2\right).$$

Goals of Computational Experiment

- Datasets: MNIST, Fashion-MNIST (60k train, 10k test)
- MLP: 2 hidden layers, $\sim 10^5$ parameters
- Subspace dimension $d = 10$, Monte Carlo samples $K = 64$, $\sigma = 1$
- Compare empirical Δ_k vs theoretical bound across k

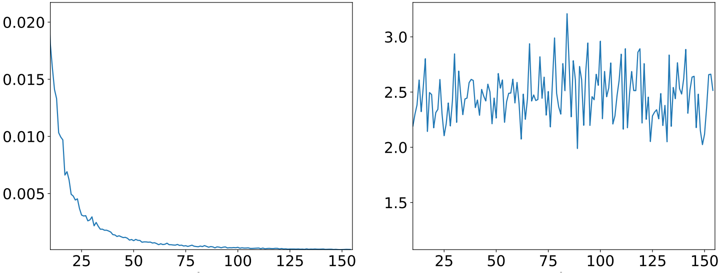
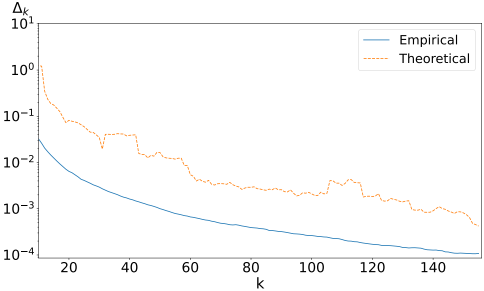


Figure 8: Monte Carlo Δ_k vs k and $\Delta_k \cdot k^2$ (p.11)

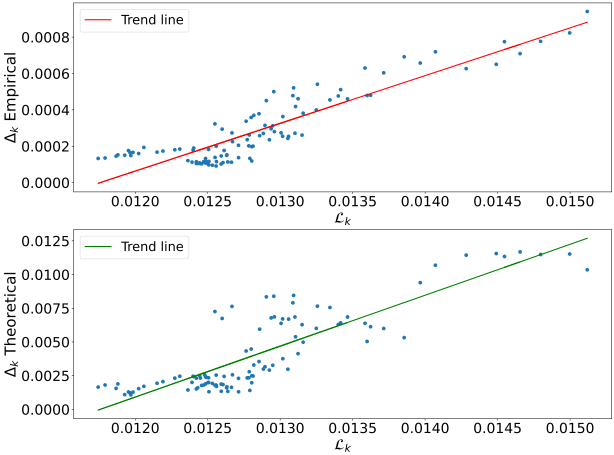
Error Analysis

- Empirical Δ_k consistently below theoretical estimate
- Gap due to neglected eigen-modes beyond top-d
- Monte Carlo variance decreases with sample size K



Theoretical (dashed) vs empirical (solid) Δ_k

Main Experiment: Theoretical vs. Empirical Δ_k



Empirical and Theoretical Δ_k

Practical Measurements

Dataset	**Model**	Δ	k	L_k	**Time (s)**
MNIST	Single-layer MLP	0.025	100	0.013	2000
	Multi-layer MLP	0.025	40	0.010	3500
	Convolutional CNN	0.025	60	0.024	1800
Fashion-MNIST	Single-layer MLP	0.030	120	0.020	2100
	Multi-layer MLP	0.030	90	0.017	4400
	Convolutional CNN	0.030	70	0.015	2400

Algorithm 1: Determine Δ-sufficient dataset size

1. Initialize dataset $D \leftarrow \emptyset$ and batch counter $k \leftarrow 0$.
2. **Repeat:**
   - $k \leftarrow k + 1$.
   - Sample a new batch of data and append to D.
   - Train or update model on D.
   - Compute top-d Hessian eigenvalues $\{\lambda_k^{(i)}\}_{i=1}^d$ and $\{\lambda_{k+1}^{(i)}\}$.
   - Estimate

$$\Delta_k \approx \frac{\sigma^4}{4}\Big(2\sum_{i=1}^{d}(\lambda_{k+1}^{(i)} - \lambda_k^{(i)})^2 + \big(\sum_{i=1}^{d}(\lambda_{k+1}^{(i)} - \lambda_k^{(i)})\big)^2\Big).$$

3. **Until** $\Delta_k < \Delta_{\text{tol}}$.
4. **Return** $|D|$, the Δ-sufficient sample size.

Results and Conclusions

- Identified dataset thresholds k^* for MNIST and Fashion-MNIST.
- Loss landscape stabilizes: additional data negligible beyond k^*.
- Hessian-based bound provides a reliable upper-bound for Δ_k.
- Proposed a practical algorithm for Δ-sufficient dataset sizing.
- Offers guidelines for dataset collection and early stopping.