

# Neural Networks Loss Landscape Convergence in Hessian Low-Dimensional Space

Tem Nikitin, Nikita Kiselev, Vladislav Meshkov, Andrey  
Grabovoy

Moscow Institute of Physics and Technology

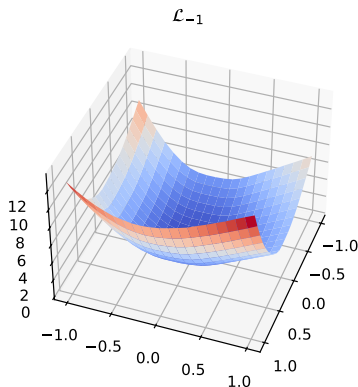
2025

# Neural Networks Loss Landscape Convergence in Hessian Low-Dimensional Space

## Goals and Tasks

1. Study how neural network loss landscape changes with dataset size
2. Define and measure metrics  $\Delta_k$
3. Derive theoretical bound for  $\Delta_k$  via top- $d$  eigenvalues
4. Propose an algorithm to determine the  $\Delta$ -sufficient dataset size

# Loss landscape stabilizes after sufficient sample size.



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  $k^*$

# Problem Statement

## Hypothesis

Beyond some  $k^*$ , adding new samples changes the local loss landscape by less than a tolerance  $\Delta_{tol}$ , 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:  $\mathbf{H}_k(\mathbf{w}) = \nabla_{\mathbf{w}}^2 \mathcal{L}_k(\mathbf{w})$

## Criteria

Convergence rate:  $\Delta_k = O(1/k^2)$

Theoretical bound via top- $d$  eigenvalues upper-bounds empirical  $\Delta_k$

Plateau in eigenvalue differences  $\lambda_i^{k+1} - \lambda_i^k$  indicates threshold

# Theoretical Analysis

Project parameters:

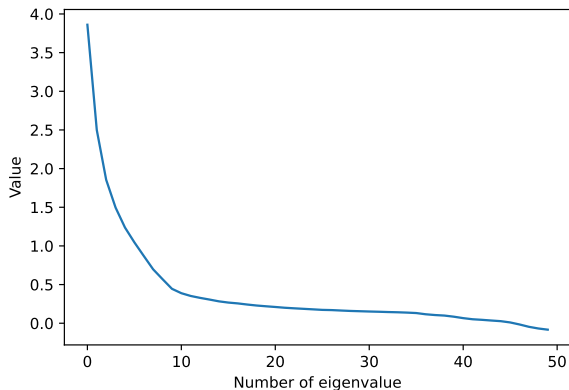
$$\mathbf{w} = \mathbf{w}^* + \mathbf{P}\boldsymbol{\theta},$$

$$\mathbf{P} = [\mathbf{e}_1, \dots, \mathbf{e}_d]$$

Taylor approx:

$$\mathcal{L}_k(\mathbf{w}^* + \mathbf{P}\boldsymbol{\theta}) \approx$$

$$\mathcal{L}_k(\mathbf{w}^*) + \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Lambda}_k \boldsymbol{\theta}$$



Eigenvalue decay

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

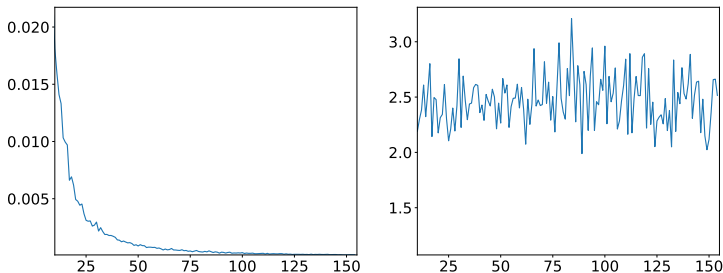
# Computational Experiment

Datasets: MNIST, Fashion-MNIST (60k train, 10k test)

MLP: 2 hidden layers,  $\sim 10^5$  parameters

Subspace dimension  $d = 10$ , Monte Carlo samples  $K = 64$ ,  
 $\sigma = 1$

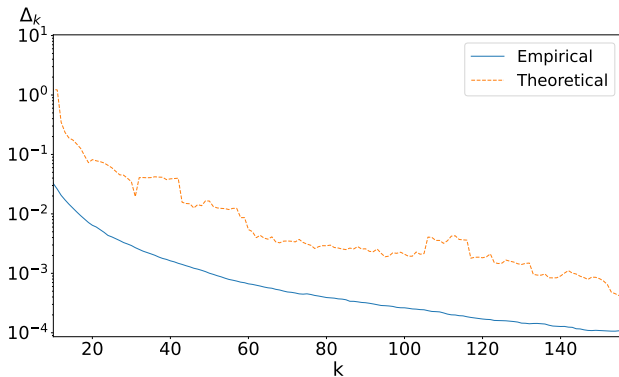
Compare empirical  $\Delta_k$  vs theoretical bound across  $k$



Monte Carlo  $\Delta_k$  vs  $k$  and  $\Delta_k \cdot k^2$

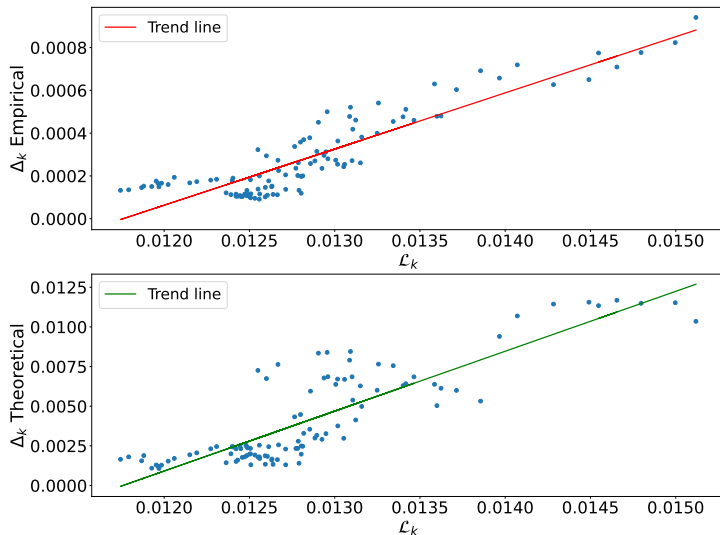
## Combining results

Empirical  $\Delta_k$  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)  $\Delta_k$

# Theoretical vs. Empirical $\Delta_k$





# Practical Measurements

Dataset	Model	$\Delta$	$k$	$L_k$	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: $\Delta$ -sufficient Dataset Size

## Algorithm 1: Determine $\Delta$ -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

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

3. **Until**  $\Delta_k < \Delta_{\text{tol}}$ .

4. **Return**  $|D|$ , the  $\Delta$ -sufficient sample size.

# Results and Conclusions

1. Identified dataset thresholds  $k^*$  for MNIST and Fashion-MNIST.
2. Loss landscape stabilizes: additional data negligible beyond  $k^*$ .
3. Hessian-based bound provides a reliable upper-bound for  $\Delta_k$ .
4. Proposed a practical algorithm for  $\Delta$ -sufficient dataset sizing.
5. Offers guidelines for dataset collection and early stopping.

1. Wu *et al.* (2017): loss landscapes vs dataset size
2. Sagun *et al.* (2018): Hessian low effective rank
3. Li *et al.* (2018): visualizing loss surfaces
4. Ghorbani *et al.* (2019): eigenvalue density analysis
5. Bousquet & Elisseeff (2002): stability and generalization bounds