EPFL

An Empirical Investigation of the Failure Mode of Training in Mildly Overparameterized NNs

Master's Thesis
- End-term Presentation-

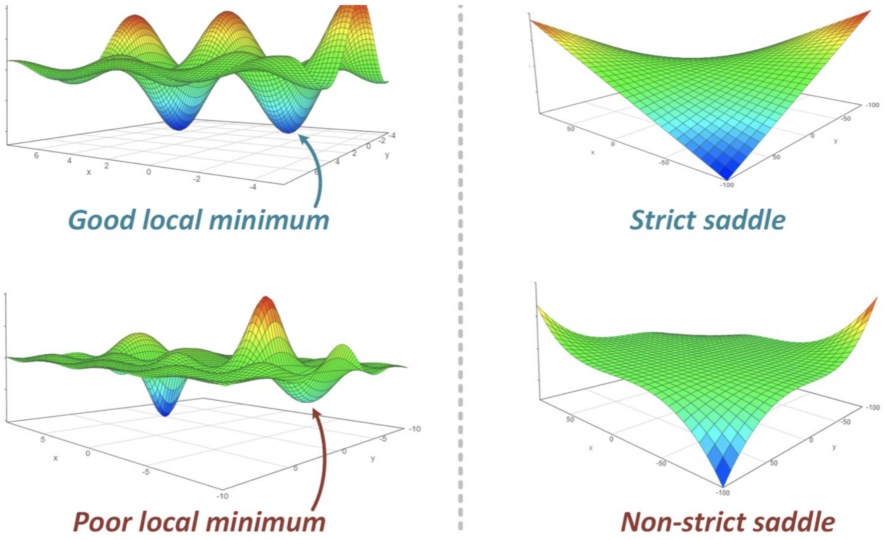
Student: *Theodor Stoican*
Supervisor: *Berfin Şimşek*

Outline

- Introducing the Problem
    - The Nature of the Landscape
    - Local Minima near Saddles
- Toy Setup
    - Overview
    - Escaping Local Minima
    - Ways of Getting to the Saddle
- MNIST Setup
    - Properties of the Saddle Line

Introducing the Problem

- Start with a neural network (normal regime)
- Add one neuron (mild OP regime) and retrain
  - What is the nature of the points at convergence?



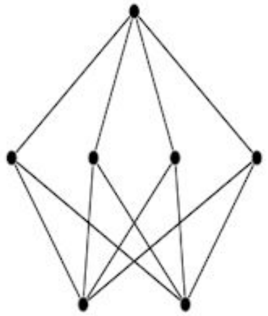
Example of neuron-points at convergence. © offconvex.org

Introducing the Problem (2)

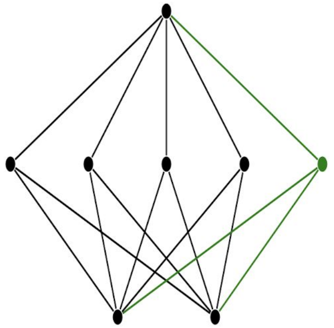
- The Nature of the Landscape - e.g. questions:
  - How common are local minima?
  - Do we ever get stuck at non-strict saddles?
- For symmetry-induced critical points ([1])
  - Is there any local minimum nearby?
  - Can we escape to it?

Toy Setup

- Teacher-student setup (same setup as the one published in [1])



4-sized teacher



A 5-sized student

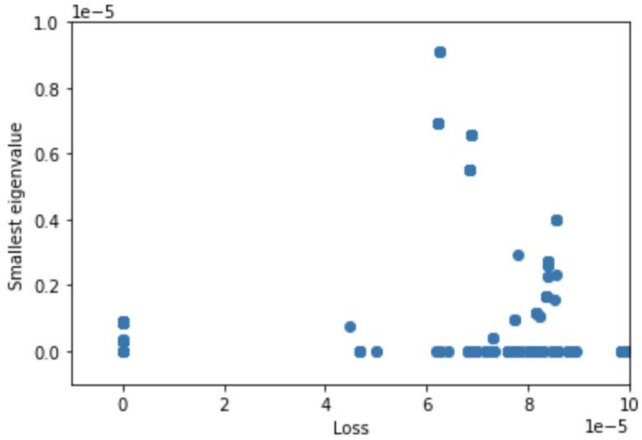
- Dataset
  - 1681 points on a regular grid $\{(x_1, x_2) | 4x_1 = -20, \ldots, 20, 4x_2 = -20, \ldots, 20\}$
  - With labels $y = \Sigma_{i=1}^{4} a_i \sigma(\Sigma_{j=1}^{2} w_{ij} x_j)$, where w_{ij} - preset weights of teacher, and $a_1 = 1, a_2 = -1, a_3 = 1, a_4 = -1$

Overview

- Out of 1000 experiments:
  - Those l.t. 1e-5 - ***global minima***
  - Those h.t. than 1e-5 - ***local minima***

		SI Critical Points	Other Local Minima
Global Minima	574	-	-
Local minima	426	55	371
Total	1000		

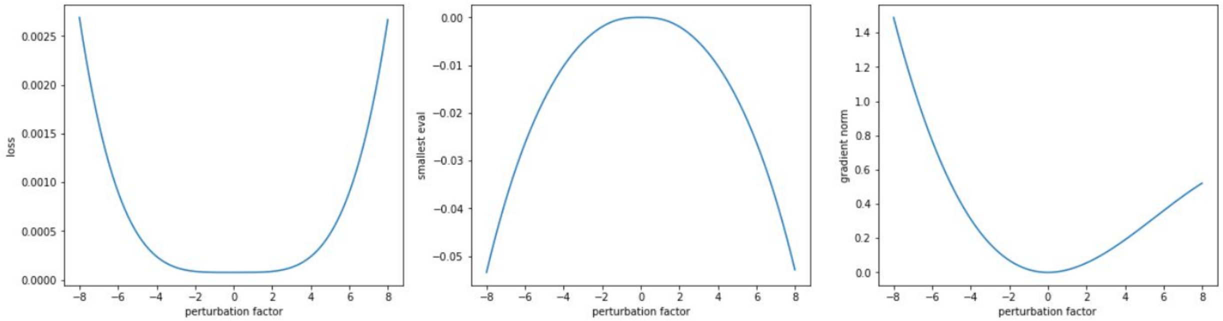
Overview of the nature of the points



Smallest eigenvalue and loss of neuron-points at convergence.

Escaping From Local Minima

- 77% of the points have a smallest eigenvalue of 0
- Any escape possible?
- **Idea**: perturb in the direction of the smallest evector
    - No escape
    - Not even by taking into account other eigenvectors

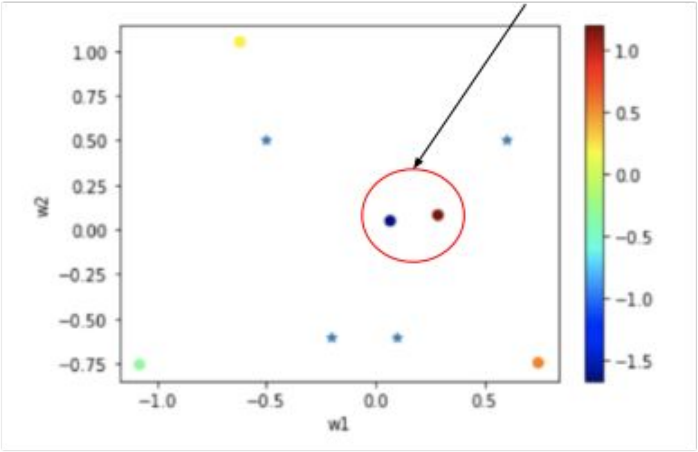


Perturbation across the smallest eigenvector for a sample failure point.

Detecting the Nearest Saddle

- For the local minima we found, we ask further:
  - Are they in the vicinity of an SI saddle?
  - **Idea: Identify the points which have a pair of neurons close to each other**

Could be merged to obtain 1 neuron.



Sample neuron-point where 2 incoming vectors are geometrically close to each other.

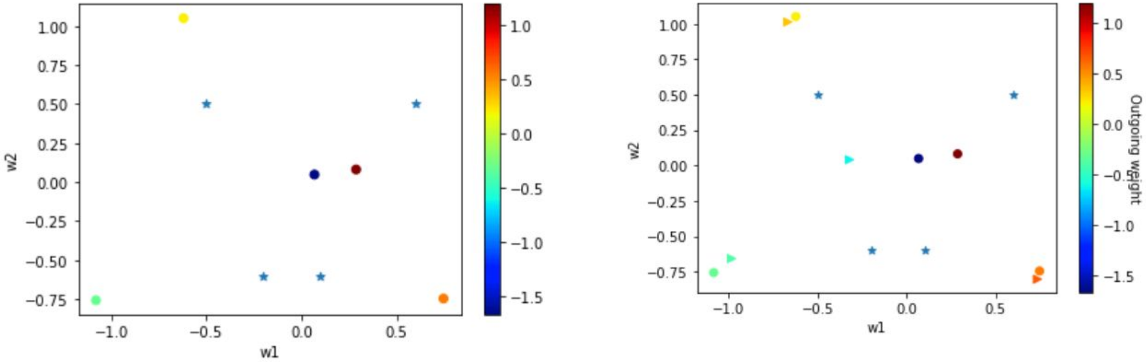
Ways of Getting to the Saddle

- Algorithm :
  - Choose the 2 closest neurons
  - Merge them
  - Obtain a reduced NN
  - Retrain the new NN from there

$$\begin{array}{ll} (w_1^1, w_2^1, a^1) \searrow & \\ & \left(\frac{w_1^1+w_1^2}{2}, \frac{w_2^1+w_2^2}{2}, a^1 + a^2\right) \\ (w_1^2, w_2^2, a^2) \nearrow & \end{array}$$

Ways of Getting to the Saddle (2)

- Example

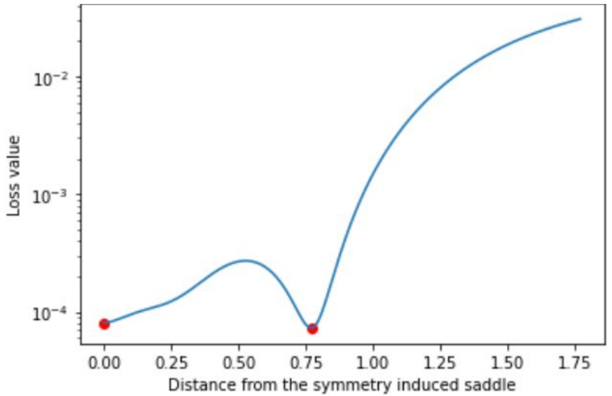


Sample neuron point, before(left) and after merging 2 of the student's neurons(right).

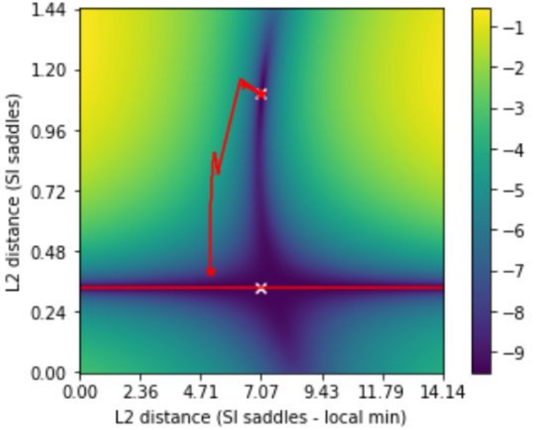
- Next:
  - Extend the reduced NN neuron-point into a saddle line
  - Find the closest point on the line to the local minimum

Distance to SI Saddle - 1D

- To check whether one such local minima (where two incoming vectors are close together) is close to an SI-saddle:
  - Evaluate the loss on the 1D line between min and closest saddle
  - View the entire landscape between the min. and the saddle line



Evolution of loss(log) across the 1D line between the closest saddle and the local min. for a sample.



Evolution of loss(log) across the 2D plane between the closest saddle line and the local min. for a sample.

Applying Theoretical Results

- Empirically
  - The local min. is within a distance of ~1.1 to the saddle line
- Formally
  - Does our distance fulfil this theorem?

Theorem 2.1. *Assume that the saddle point x^* has index-1: its Hessian has one negative eigenvalue and the other eigenvalues are positive. For all unit u such that $u^T \nabla^2 f(x^*) u \leq 0$, let $B = \min_{\|u\|=1} \max\{\nabla^3 f(x^*)(u), \nabla^3 f(x^*)(-u)\}$. We assume $B > 0$ exists. Assume that there is an $R > 0$ such that $\nabla^4 f(\xi)(u) \geq 0$ for all $\|\xi - x^*\| < R$ and $\|u\| = 1$. If*

$$\frac{-3\lambda_{\min}}{B} \leq R,$$

then we have a local min. within an l_2-distance $-3\lambda_{\min}/B$ from the saddle point x^*.

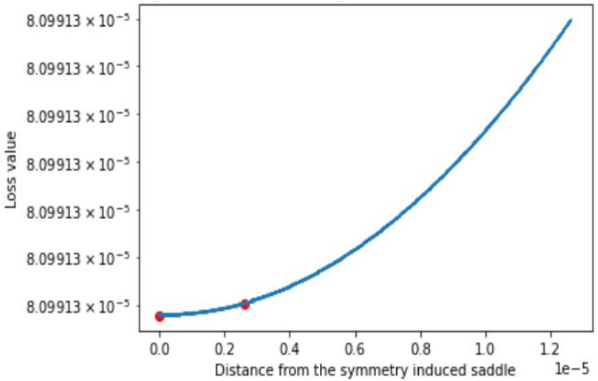
EPFL

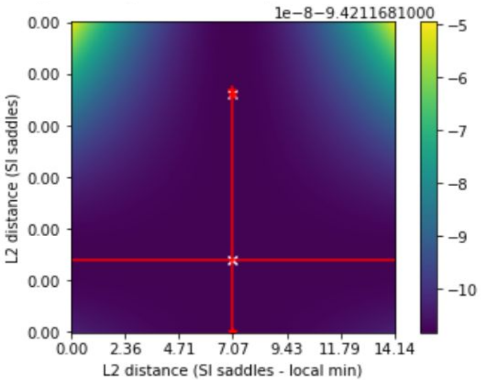
Applying Theoretical Results (2)

- The conditions of the theorem are not fulfilled
  - In particular, the fourth derivative is not positive for every ξ in the theorem
- Hence:
  - The closest saddle we find may not really be "the closest"
  - The theorem's precondition may be relaxed for such cases

SI Critical Points as Failure Points

- Furthermore, some of the local minima seem to be SI critical points
  - The distance between the closest saddle and the min. is ~0





Example of the small distance between the local min. and the identified closest saddle (1e-5 - very close to 0) .

- Can we formally investigate their nature?

SI Critical Points as Failure Points (2)

- We can use the following conditions:

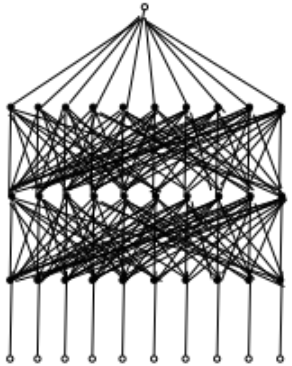
Conditions for no negative eigenvalue: If $U_{ij} = 0$ for all $j \in [d]$,$i \in [d_0]$ and $\mu(1-\mu)Y$ has no negative eigenvalues, then the min. eigenvalue of the Hessian at this critical point $(w^*, \mu a^*, w^*, (1-\mu)a^*)$ is 0.

1. $Y = \hat{\mathbb{E}}[\sigma''(w^* \cdot x)a^* \cdot e(x)xx^T] \in \mathbb{R}^{d \times d}$ has at least one negative and at least one positive eigenvalue,
2. $U_{ij} = \hat{\mathbb{E}}[\sigma'(w^* \cdot x)e(x)_i x_j] \neq 0$ for some $j \in [d]$, $i \in [d_0]$.

- For one such neuron-point
  - Find its origin in the reduced NN
  - Evaluate U_{ij}
  - Evaluate the eigenvalues of the $\mu(1-\mu)Y$ matrix
- **~50 points (13%)** of failures are SI critical points

MNIST Setup

- How does the theory apply to a more world-like scenario?
- In particular:
  - A fully-connected 3-layer NN
  - Dataset
    - Inputs: top 10 PCA components
    - Labels: **1** for odd, **-1** for even
  - MSE Loss



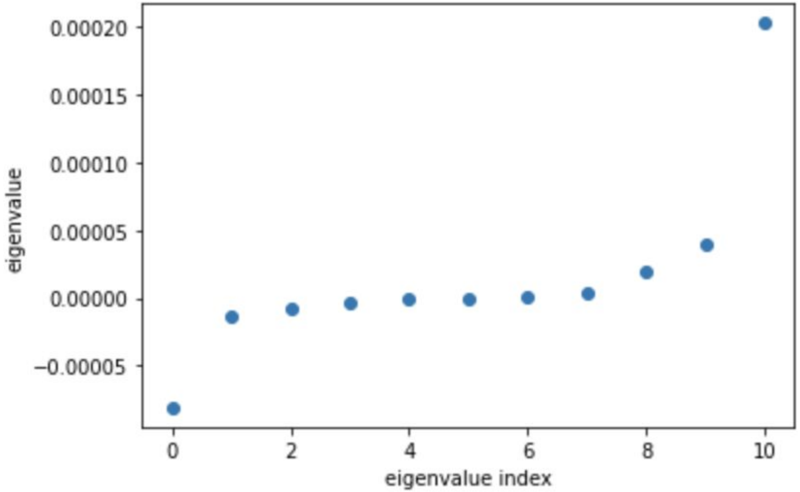
3-layer fully-connected NN with 10 neurons on each hidden layer. The output layer has 1 neuron only.

MNIST Setup (2)

- Procedure:
  - Train and find a local min. with this new setup
    - Stop condition: grad = ~1e-6
  - Duplicate a neuron on a layer
    - **(w, a) -> (w, μ * a), (w, (1- μ) * a)**
    - Vary μ
  - Perturb and retrain until a local min. is reached

Saddle Line - Verification

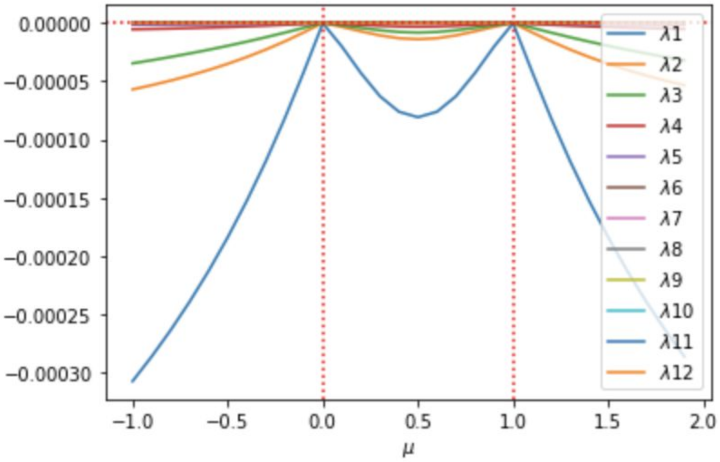
- Make sure the local min. we find is a failure mode in the overparameterized NN
  - Check the **Y** and **U** matrices
    - All $U_{ij} = 0$ as before (only 1 output neuron)
    - Y must have neg. and pos. eigenvalues



The eigenspectrum of the Y matrix.

Saddle Line - Duplicating on the last layer

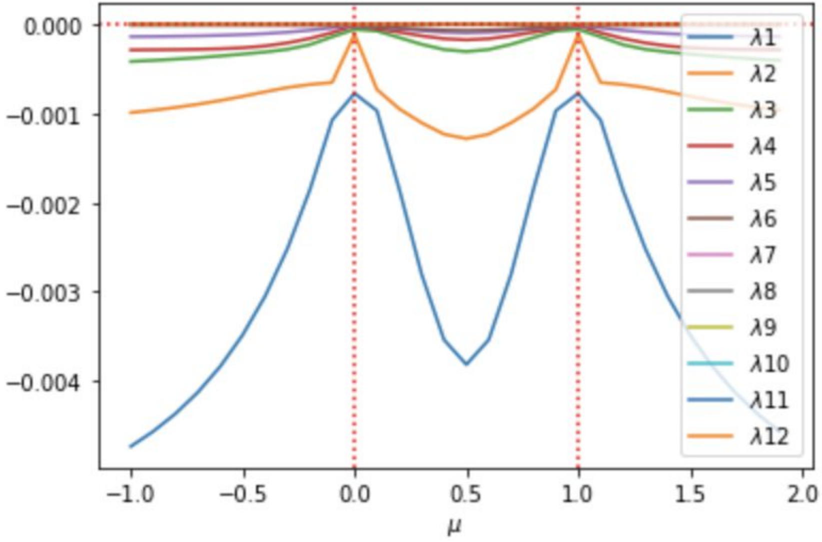
- Furthermore, inspect the eigenvalues (of H) across the saddle line
- 12 eigenvalues are expected to cross 0 at $\mu \in \{0, 1\}$
  - Intuition: the Hessian will have duplicate rows after duplicating a neuron



The evolution of the 12 smallest eigenvalues on the SI saddle for a range of μ

Saddle Line - Duplicating on the second layer

- 4 eigenvalues cross 0 at $\mu = 0$



The evolution of the 12 smallest eigenvalues on the SI saddle for a range of μ

Finding the Closest Local Min.

- Idea:
  - Vary **μ**
  - Perturb (with an isotropic Gaussian)
  - Train until the gradient is small (~1e-6)
- For $-1 < \mu < 2$ the algorithm always finds a global minimum after perturbation
- **TODO**:
  - Investigate potential convergence issues - why a global min. ?
  - Experiment with other seeds

Q&A

References

[1] Şimşek Berfin et al., 2021, Geometry of the Loss Landscape in Overparameterized Neural Networks: Symmetries and Invariances

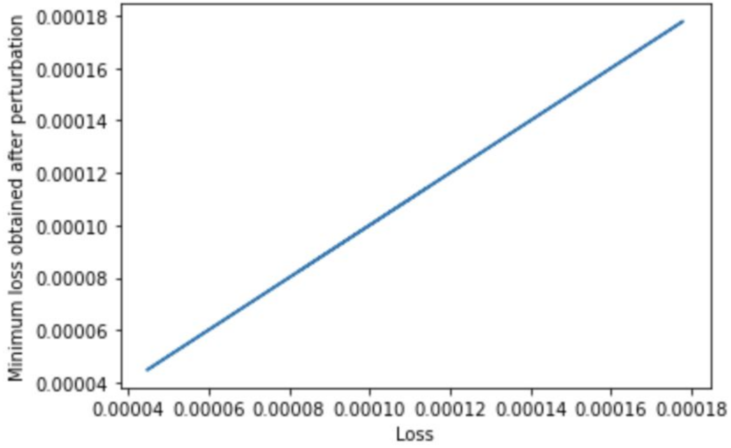
[2] Itay M Safran et al., 2021, The Effects of Mild Over-parameterization on the Optimization Landscape of Shallow ReLU Neural Networks

[3] Brea Johanni et al., 2019, Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape

[4] Zhang Yaoyu, 2021, Embedding Principle of Loss Landscape of Deep Neural Networks

Escaping From Local Minima (2)

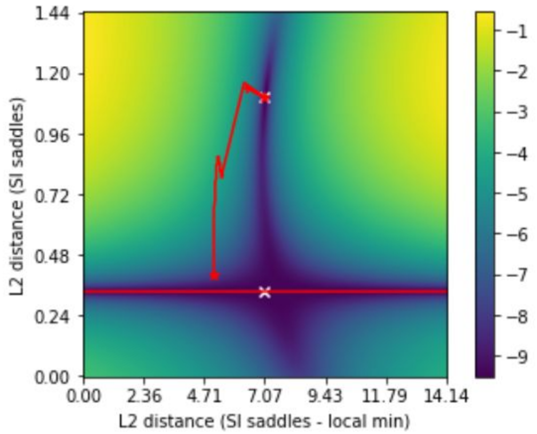
- Still, by perturbing, the smallest eigenvector can change direction during perturbation
- We investigate this, by adapting our algorithm:
  - By finding all 0 eigenvalue directions at any point during perturb.
  - By testing all these directions recursively
- Even so, for no point do we manage to escape



Relationship between the loss at the failure point and the min. loss obtained after perturbation.

Distance to SI Saddle - 2D

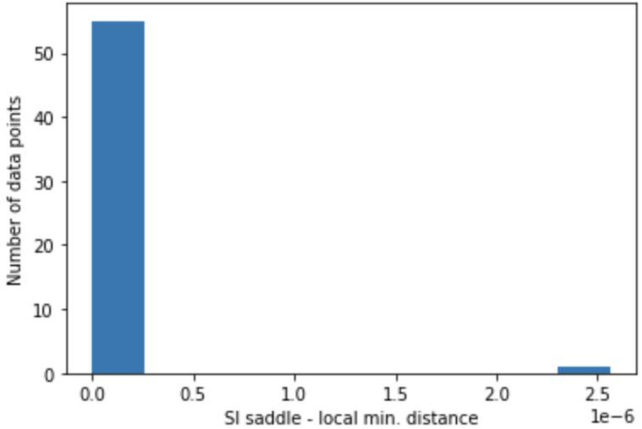
- For a better picture:
  - View the entire landscape between the min. and the saddle line
  - Project the GD trajectories on this plane



Evolution of loss(log) across the 2D plane between the closest saddle line and the local min. for a sample.

SI Critical Points as Failure Points (3)

- Evaluating U_{ij}
  - $L = \sum_i \sigma(w_i \cdot x) \cdot a_i$
  - At the saddle:
    - $\frac{\partial L}{\partial w_i} = x \cdot \sigma'(w_i \cdot x) \cdot a_i = 0 \Rightarrow \sigma'(w_i \cdot x) = 0$
    - Hence, all U_{ij} are 0
- Evaluating $\mu(1-\mu)Y$
  - No negative eigenvalues for **~13%** of the local minima
  - Hence, ~50 points of failure are SI critical points.



The distribution of the distances (local min. - SI saddle) for the local min. which are SI critical points.