Neural Network Landscape

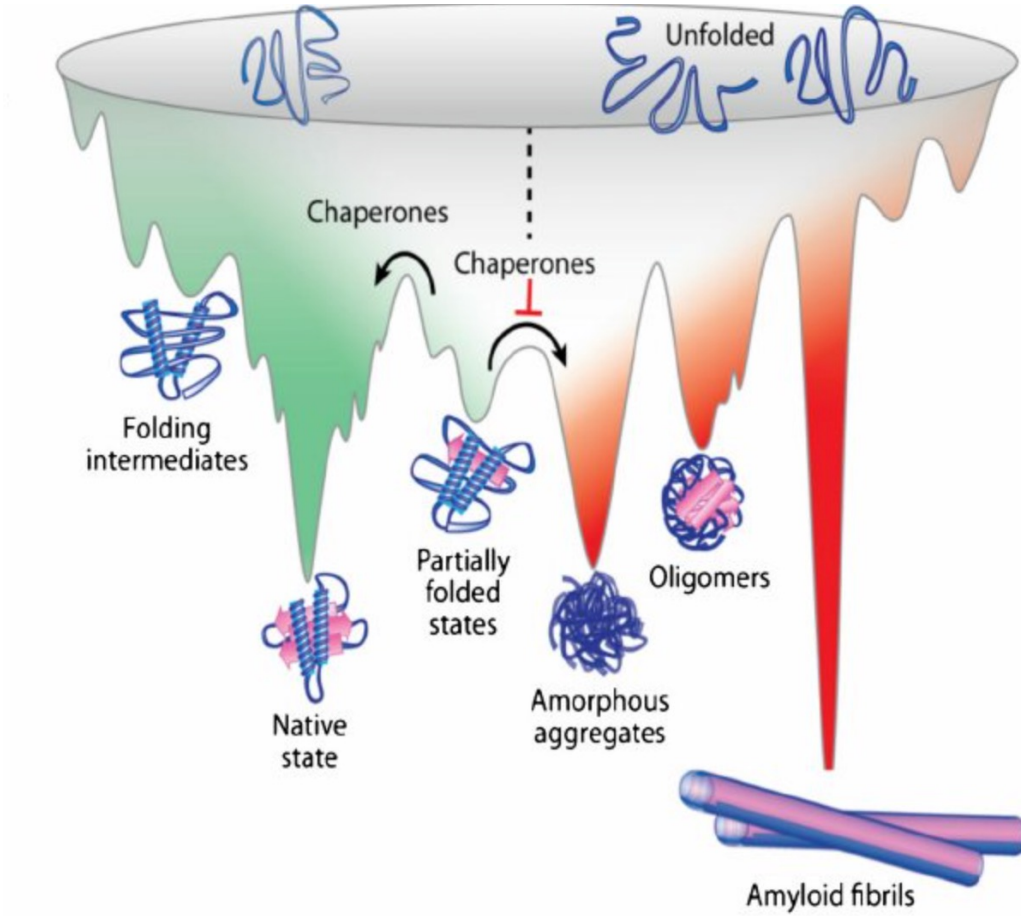
Lei Wu

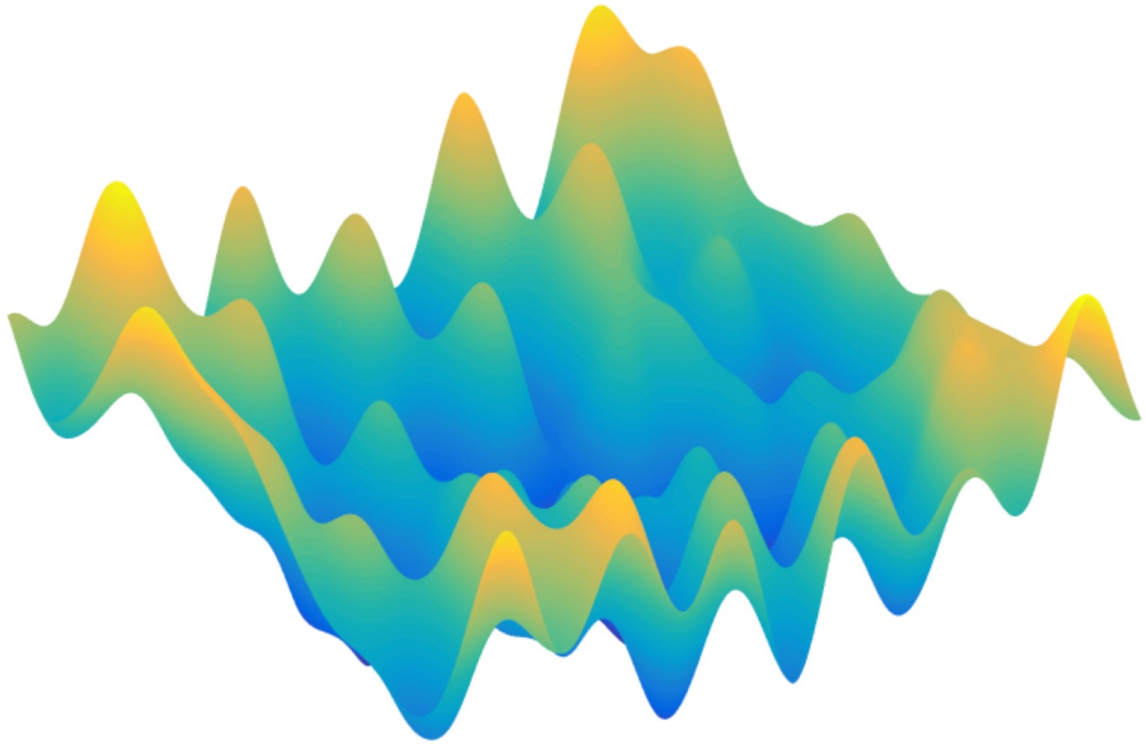
Big Picture

- Landscape properties are essential for understanding the optimization, generalization and implicit bias of neural network training.
- We have seen that sharpness of local landscape is related to generalization performance and implicit bias.
- Progressive sharpening impacts the convergence speed of optimizers.
- In this lecture, we shall mostly focus on exploring the **non-local properties** of neural network landscape.

Mountain Landscape

Non-convex Landscape in Science





The loss landscape of neural network

- In general, the loss landscape of neural network can be also extremely bad. There are many papers arguing this in 1990s.
- This is not surprising as the landscape property highly depends on the target function.

Exponentially many local minima for single neurons

Peter Auer **Mark Herbster** **Manfred K. Warmuth**

Department of Computer Science
Santa Cruz, California
{pauer,mark,manfred}@cs.ucsc.edu

Abstract

We show that for a single neuron with the logistic function as the transfer function the number of local minima of the error function based on the square loss can grow exponentially in the dimension.

Recap of Auer et al., 1995

- Consider the learning of a single neuron with sigmodal activation function:

$$E_S(W) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sigma(w^\top x_i))^2.$$

Theorem 3.4 *Let ϕ and L satisfy (P1). Then for all $n \geq 1$ there is a sequence of examples $\mathcal{S} = \langle(\mathbf{x}_1, y), \ldots, (\mathbf{x}_n, y)\rangle$, $\mathbf{x}_t \in \mathbf{R}^d$, $y \in \phi(\mathbf{R})$, such that $E_{\mathcal{S}}(\mathbf{w})$ has $\left\lfloor \frac{n}{d} \right\rfloor^d$ distinct local minima.*

Proof Sketch

- First, prove the for d=1, it holds
  - By induction, assume there exists $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ such that there exist n local minima. Then, show we can construct a(x_{n+1}, y_{n+1}) such that $E_{S'}(\cdot)$ has n+1 minima, where $S' = \{(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\}$.
- Second, lift to high dimension by the following lemma.

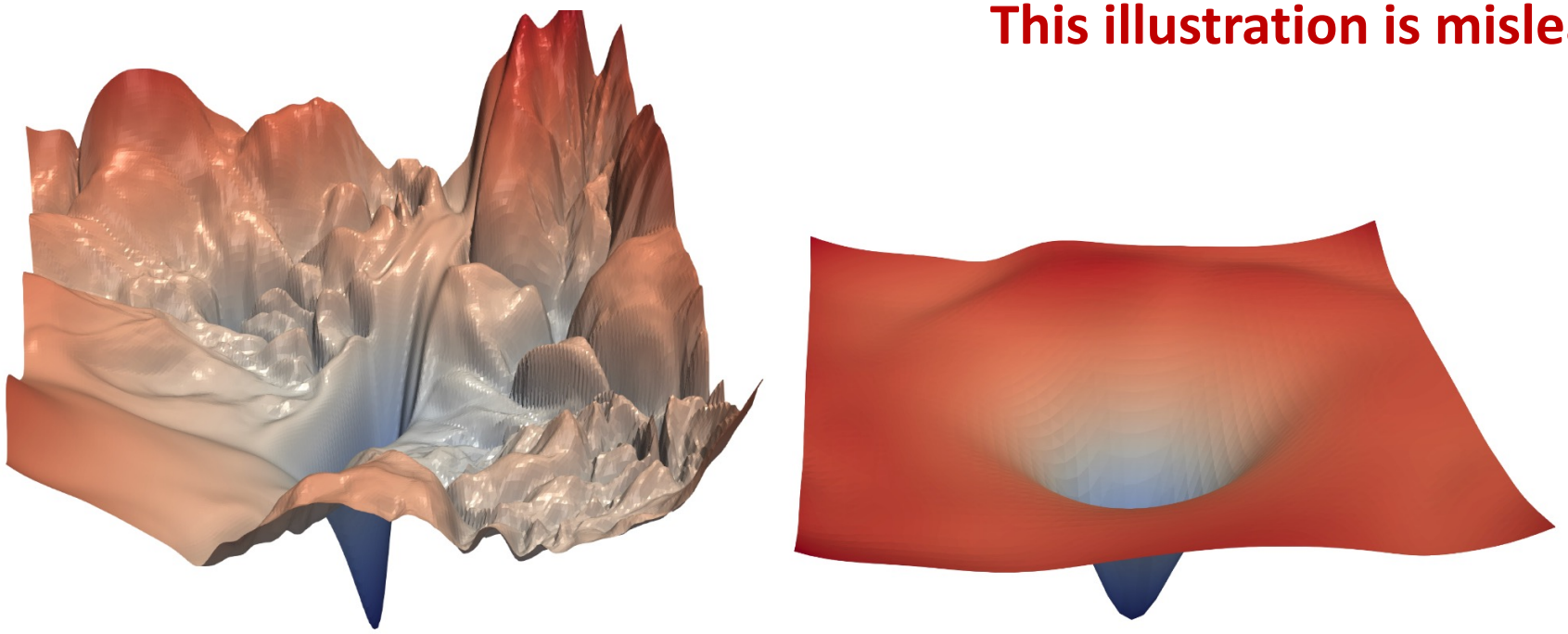
Lemma 3.3 *Let $f : \mathbf{R} \to \mathbf{R}$ be a continuous function with n disjoint minimum-containing sets $U_1, \ldots, U_n$. Then the sets $U_{t_1} \times \cdots \times U_{t_d}$, $t_j \in \{1, \ldots, n\}$, are n^d disjoint minimum-containing sets for the function $g : \mathbf{R}^d \to \mathbf{R}$, $g(x_1, \ldots, x_d) = f(x_1) + \cdots + f(x_d)$.*

- Third, construct $S = \cup_{k \in [d]} S_k$, where

$$S_1 = \{((x_i, 0, \ldots, 0), y_i)\}_{i=1}^{n/k}, \quad S_2 = \{((0, x_i, 0, \ldots, 0), y_i)\}_{i=n/k+1}^{2n/k}, \cdots$$

A Modern View of Neural Network Landscape

This illustration is misleading.

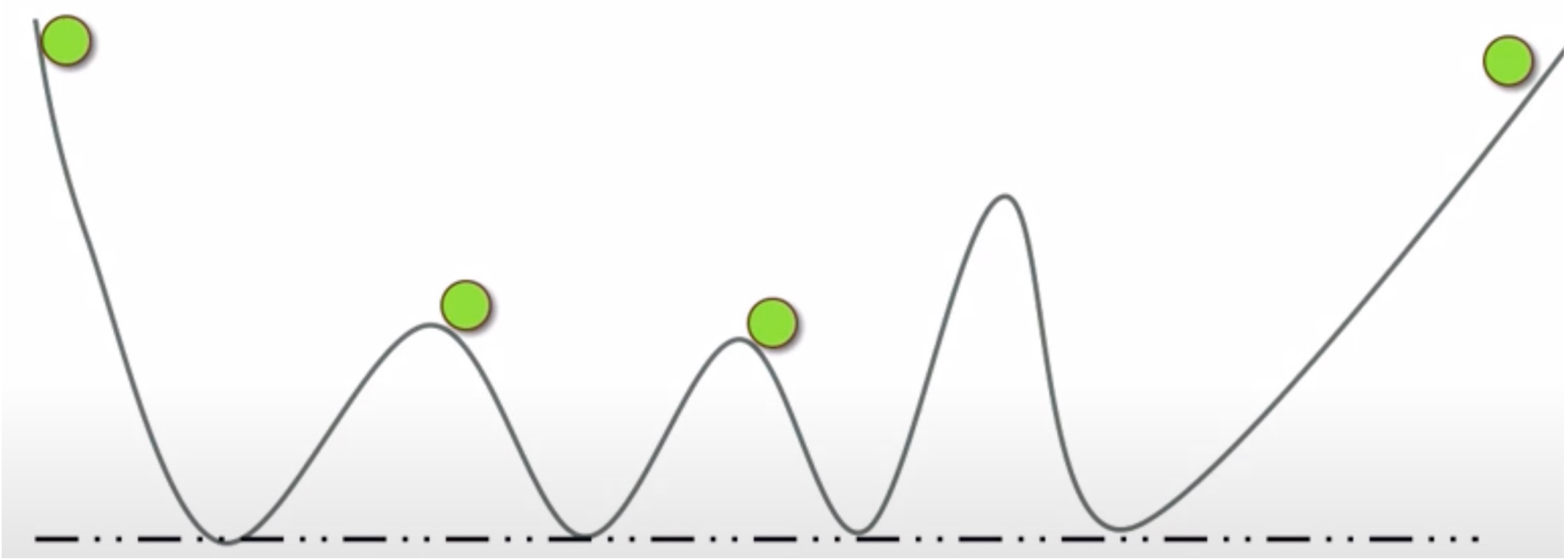


(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

Taken from (Li et al., NIPS 2017) .

Absent of Spurious Local Minima



Linear Networks (Kenji Kawaguchi, NIPS2016)

Theorem 2.3 (Loss surface of *deep* linear networks) *Assume that* XX^T *and* XY^T *are of full rank with* $d_y \leq d_x$ *and* Σ *has* d_y *distinct eigenvalues. Then, for any depth* $H \geq 1$ *and for any layer widths and any input-output dimensions* $d_y, d_H, d_{H-1}, \ldots, d_1, d_x \geq 1$ *(the widths can arbitrarily differ from each other and from* d_y *and* d_x*), the loss function* $\bar{\mathcal{L}}(W)$ *has the following properties:*

(i) It is non-convex and non-concave.

(ii) Every local minimum is a global minimum.

(iii) Every critical point that is not a global minimum is a saddle point.

(iv) If $\operatorname{rank}(W_H \cdots W_2) = p$*, then the Hessian at any saddle point has at least one (strictly) negative eigenvalue.*[1]

$$X \in \mathbb{R}^{d_x \times m}, Y \in \mathbb{R}^{d_y \times m}, \Sigma = YX^T(XX^\top)^{-1}XY^T, p = \min(d_1, d_2, \ldots, d_H).$$

NTK results

- For empirical landscape, as long as the network is sufficiently large, there exist many global minima near the initialization and GD can find them efficiently.
- It seems to suggest that overparameterization is a key to ensuring a benign landscape
- Since then, many studies have aimed to establish certain benign properties of neural network landscape by exploiting over-parameterization.

Analysis of the teacher-student setup

- Teacher-student setup

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_k} \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\frac{1}{2}\left(\sum_{i=1}^{n}[\mathbf{w}_i^\top\mathbf{x}]_+ - \sum_{i=1}^{k}[\mathbf{v}_i^\top\mathbf{x}]_+\right)^2\right]$$

- When n=k, there exists many local minima and SGD can find them easily.
- When n=k+1, SGD +rand init rarely find local minima
- When n>=k+2, SGD+rand init nearly cannot find local minima

Itay Safran, Ohad Shamir, Spurious Local Minima are Common in Two-Layer ReLU Neural Networks, ICML 2018.

Experiments

Table 1. Spurious local minima found for $n = k$

k	n	% of runs converging to local minima	Average minimal eigenvalue	Average objective value
6	6	0.3%	0.0047	0.025
7	7	5.5%	0.014	0.023
8	8	12.6%	0.021	0.021
9	9	21.8%	0.027	0.02
10	10	34.6%	0.03	0.022
11	11	45.5%	0.034	0.022
12	12	58.5%	0.035	0.021
13	13	73%	0.037	0.022
14	14	73.6%	0.038	0.023
15	15	80.3%	0.038	0.024
16	16	85.1%	0.038	0.027
17	17	89.7%	0.039	0.027
18	18	90%	0.039	0.029
19	19	93.4%	0.038	0.031
20	20	94%	0.038	0.033

Table 2. Spurious local minima found for $n \neq k$

k	n	% of runs converging to local minima	Average minimal eigenvalue	Average objective value
8	9	0.1%	0.0059	0.021
10	11	0.1%	0.0057	0.018
11	12	0.1%	0.0056	0.017
12	13	0.3%	0.0054	0.016
13	14	1.5%	0.0015	0.038
14	15	5.5%	0.002	0.033
15	16	10.1%	0.004	0.032
16	17	18%	0.0055	0.031
17	18	20.9%	0.007	0.031
18	19	36.9%	0.0064	0.028
19	20	49.1%	0.0077	0.027

Mildly over-parameterized

- Karhadkar et al., Mildly Overparameterized ReLU Networks Have a Favorable Loss Landscape, arXiv:2305.19510, 2023.
- Zhou et al., A Local Convergence Theory for Mildly Over-Parameterized Two-Layer Neural Network, COLT 2021.
- Safran et al., The Effects of Mild Over-parameterization on the Optimization Landscape of Shallow ReLU Neural Networks, COLT 2021.

Mode Connectivity

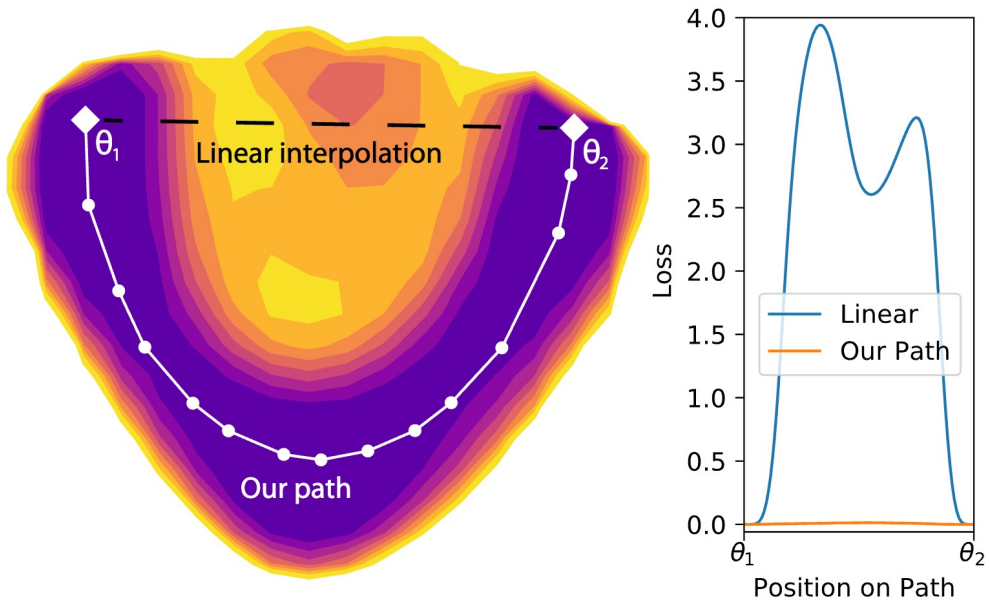
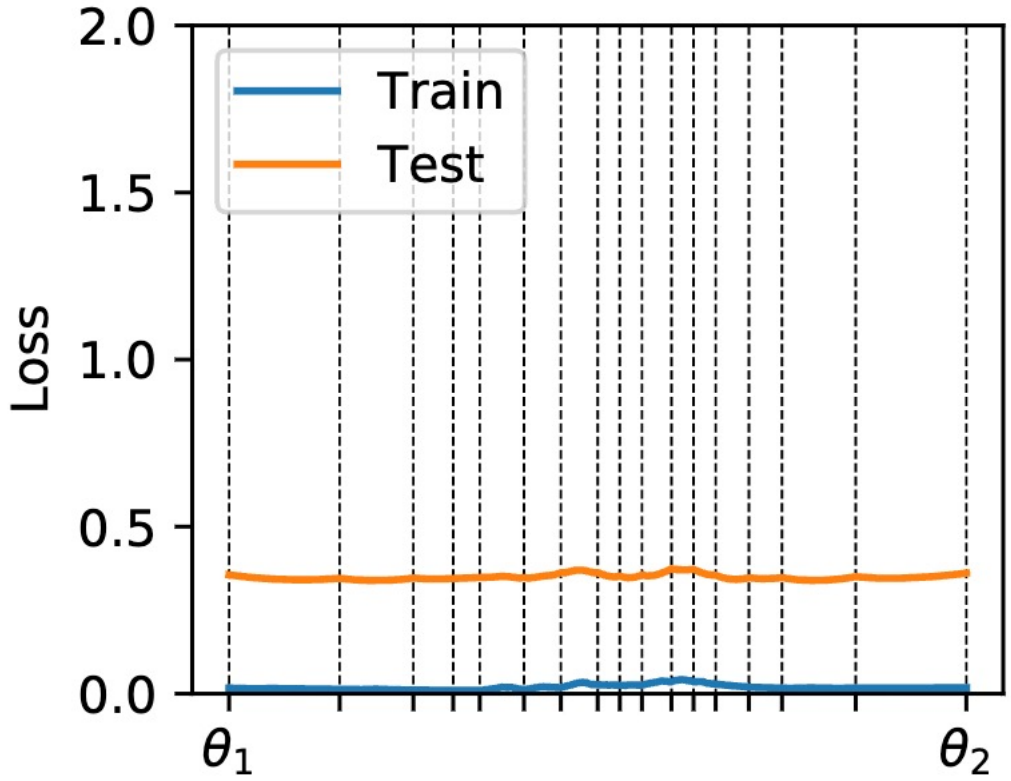
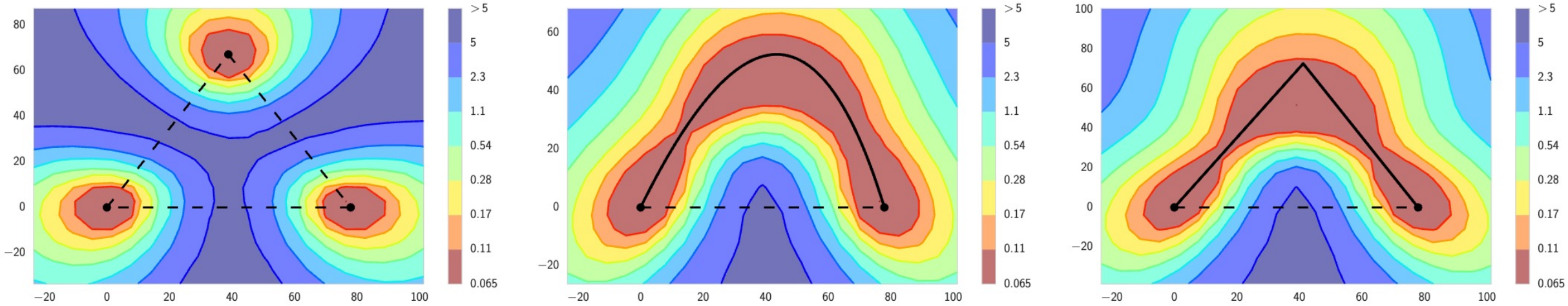


Figure 1. Left: A slice through the one million-dimensional training loss function of DenseNet-40-12 on CIFAR10 and the minimum energy path found by our method. The plane is spanned



Draxler et al., Essentially No Barriers in Neural Network Energy Landscape, ICML 2018.

Visualizing Mode Connectivity (link)



Loss surface of ResNet-164 on CIFAR-100. **Left**: three optima for independently trained networks; **Middle** and **Right**: A quadratic Bezier curve, and a polygonal chain with one bend, connecting the lower two optima on the left panel along a path of near-constant loss.

Garipov, et al., Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs, NeuIPS 2018.

Summary

- Local Landscape/geometry
  - Sharpness
  - Saddle points
  - Plateau, basin, valley
- Non-local landscape
  - Progressive sharpening (landscape along optimization trajectory)
  - Mode connectivity
- Global landscape
  - Absent of bad local minima
- Over-parametrization

Reading

- M. Bianchini and M. Gori, Optimal learning in artificial neural networks: A review of theoretical results, Neurocomputing, 1996. [Old survey]
- Ruoyu Sun et al., The Global Landscape of Neural Networks: An Overview. IEEE Signal Processing Magazine 2020. [Modern survey]
- Draxler et al., Essentially No Barriers in Neural Network Energy Landscape, ICML 2018.
- Hao Li et al., Visualizing the Loss Landscape of Neural Nets, NIPS 2017
- https://izmailovpavel.github.io/curves_blogpost/
- https://losslandscape.com/