

# Mildly Overparametrized ReLU Networks Have a Favorable Loss Landscape

Guido Montúfar  
UCLA & MPI MiS

Joint work with

Kedar Karhadkar



Michael Murray



Hanna Tseran

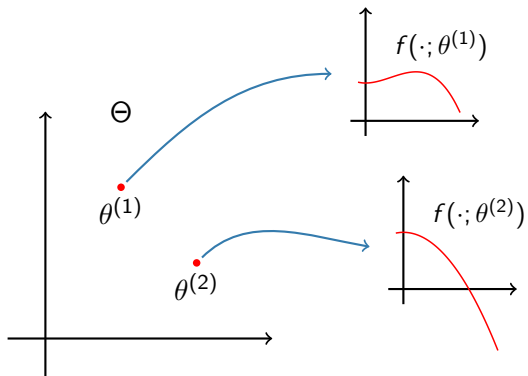


Geometry and Machine Learning, MPI MiS, Nov 2023

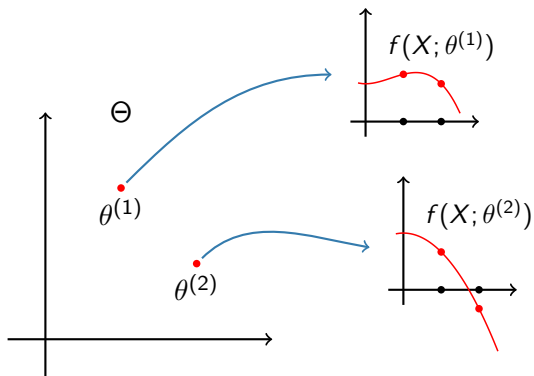


Max Planck Institute for  
**Mathematics**  
in the **Sciences**

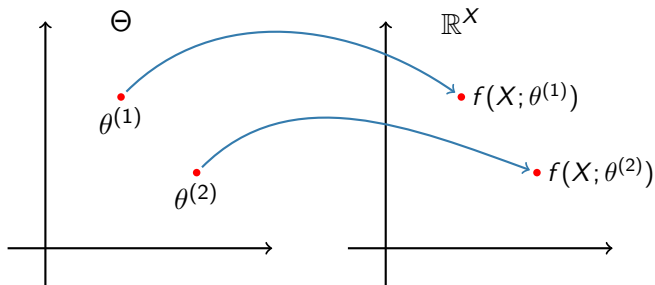




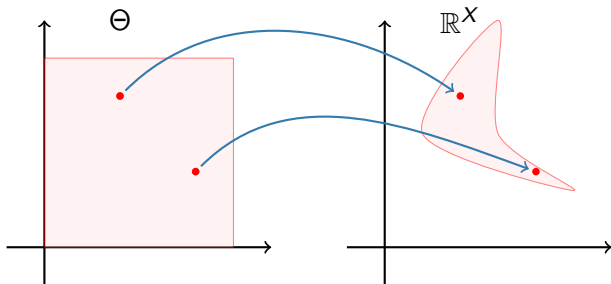
Parametric model



Parametric model and input data set



Parametric model over the input data set



Function space over the input data set

# Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus [SS89, AHW95, FA00, SCP17, SS18].

# Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus [SS89, AHW95, FA00, SCP17, SS18].
- A particularly **puzzling question** is why bad local minima do not seem to be a problem for training.

## Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus [SS89, AHW95, FA00, SCP17, SS18].
- A particularly **puzzling question** is why bad local minima do not seem to be a problem for training.
- Very **highly overparameterized** networks with  $d_1 = \Omega(n^2)$  are known to have more benevolent loss landscape and follow lazy training.

# Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus [SS89, AHW95, FA00, SCP17, SS18].
- A particularly **puzzling question** is why bad local minima do not seem to be a problem for training.
- Very **highly overparameterized** networks with  $d_1 = \Omega(n^2)$  are known to have more benevolent loss landscape and follow lazy training.
- We are able to avoid excessive overparameterization by emphasizing *qualitative* aspects of the loss landscape, using only the rank of the Jacobian rather than e.g. the smallest eigenvalue of the NTK.

## Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus [SS89, AHW95, FA00, SCP17, SS18].
- A particularly **puzzling question** is why bad local minima do not seem to be a problem for training.
- Very **highly overparameterized** networks with  $d_1 = \Omega(n^2)$  are known to have more benevolent loss landscape and follow lazy training.
- We are able to avoid excessive overparameterization by emphasizing *qualitative* aspects of the loss landscape, using only the rank of the Jacobian rather than e.g. the smallest eigenvalue of the NTK.
- We obtain theorems under more realistic **mild overparameterization**  $d_1 = \Omega(n \log n)$  or even  $d_1 = \Omega(1)$  for high-dimensional inputs.

## Overview

For  $n$  data points,  $d_0$  input dimension,  $d_1$  hidden units, we show:

## Overview

For  $n$  data points,  $d_0$  input dimension,  $d_1$  hidden units, we show:

- Theorem 2: If  $d_0 d_1 \geq n$  and  $d_1 = \Omega(\log(\frac{n}{\epsilon d_0}))$ , then all activation regions, except for an  $\epsilon$  fraction, have no bad local minima.

## Overview

For  $n$  data points,  $d_0$  input dimension,  $d_1$  hidden units, we show:

- Theorem 2: If  $d_0 d_1 \geq n$  and  $d_1 = \Omega(\log(\frac{n}{\epsilon d_0}))$ , then all activation regions, except for an  $\epsilon$  fraction, have no bad local minima.
- Theorem 9: If  $d_0 = 1$  and  $d_1 = \Omega(n \log(\frac{n}{\epsilon}))$ , all but at most an  $\epsilon$  fraction of *non-empty* activation regions have no bad local minima.

## Overview

For  $n$  data points,  $d_0$  input dimension,  $d_1$  hidden units, we show:

- Theorem 2: If  $d_0 d_1 \geq n$  and  $d_1 = \Omega(\log(\frac{n}{\epsilon d_0}))$ , then all activation regions, except for an  $\epsilon$  fraction, have no bad local minima.
- Theorem 9: If  $d_0 = 1$  and  $d_1 = \Omega(n \log(\frac{n}{\epsilon}))$ , all but at most an  $\epsilon$  fraction of *non-empty* activation regions have no bad local minima.
- Theorem 11: If  $d_0 = 1$  and  $d_1 = d_+ + d_-$  with  $d_+, d_- = \Omega(n \log(\frac{n}{\epsilon}))$ , then all but at most an  $\epsilon$  fraction of non-empty activation regions contain an affine set of global minima of codimension  $n$ .

## Setup

- We consider input and output **data**

$$X = (x^{(1)}, \dots, x^{(n)}) \in \mathbb{R}^{d \times n}, \quad y = (y^{(1)}, \dots, y^{(n)}) \in \mathbb{R}^{1 \times n}.$$

## Setup

- We consider input and output **data**

$$X = (x^{(1)}, \dots, x^{(n)}) \in \mathbb{R}^{d \times n}, \quad y = (y^{(1)}, \dots, y^{(n)}) \in \mathbb{R}^{1 \times n}.$$

- We consider a **parameterized model**

$$F: \underset{\text{parameter}}{\mathbb{R}^m} \times \underset{\text{input}}{\mathbb{R}^d} \rightarrow \underset{\text{prediction}}{\mathbb{R}}$$

and the vector of predictions on input data  $X$ ,

$$F(\theta, X) := (F(\theta, x^{(1)}), F(\theta, x^{(2)}), \dots, F(\theta, x^{(n)})).$$

## Setup

- We consider input and output **data**

$$X = (x^{(1)}, \dots, x^{(n)}) \in \mathbb{R}^{d \times n}, \quad y = (y^{(1)}, \dots, y^{(n)}) \in \mathbb{R}^{1 \times n}.$$

- We consider a **parameterized model**

$$F: \underset{\text{parameter}}{\mathbb{R}^m} \times \underset{\text{input}}{\mathbb{R}^d} \rightarrow \underset{\text{prediction}}{\mathbb{R}}$$

and the vector of predictions on input data  $X$ ,

$$F(\theta, X) := (F(\theta, x^{(1)}), F(\theta, x^{(2)}), \dots, F(\theta, x^{(n)})).$$

- The **mean squared error** loss  $L: \underset{\text{parameter}}{\mathbb{R}^m} \times \underset{\text{inputs}}{\mathbb{R}^{d \times n}} \times \underset{\text{outputs}}{\mathbb{R}^{1 \times n}} \rightarrow \mathbb{R}^1$ ,

$$L(\theta, X, y) := \frac{1}{2} \sum_{i=1}^n (F(\theta, x^{(i)}) - y^{(i)})^2. \quad (1)$$

### Lemma 1 (Full rank Jacobian implies critical point is global min)

*Fix a dataset  $(X, y) \in \mathbb{R}^{d \times n} \times \mathbb{R}^{1 \times n}$ , a parametrized model  $F$ , and a differentiable critical point  $\theta \in \mathbb{R}^m$  of the squared error loss (1).*

*If  $\text{rank}(\nabla_{\theta} F(\theta, X)) = n$ , then  $\theta$  is a global minimizer.*

### Lemma 1 (Full rank Jacobian implies critical point is global min)

Fix a dataset  $(X, y) \in \mathbb{R}^{d \times n} \times \mathbb{R}^{1 \times n}$ , a parametrized model  $F$ , and a differentiable critical point  $\theta \in \mathbb{R}^m$  of the squared error loss (1).

If  $\text{rank}(\nabla_{\theta} F(\theta, X)) = n$ , then  $\theta$  is a global minimizer.

Proof.

$$0 = \nabla_{\theta} L(\theta, X, y) = \underbrace{\nabla_{\theta} F(\theta, X)}_{\text{rank}=n} \cdot \underbrace{(F(\theta, X) - y)}_{=0}.$$



## Shallow ReLU network

- Consider **two-layer ReLU network**  $F: \underset{\text{parameter}}{\mathbb{R}^{d_1 \times d_0}} \times \underset{\text{input}}{\mathbb{R}^{d_0}} \rightarrow \underset{\text{prediction}}{\mathbb{R}}$

$$F(W, x) = v^T \sigma(Wx),$$

where  $\sigma: s \mapsto \max\{0, s\}$  componentwise, and  $v \in \mathbb{R}^{d_1}$ .

## Shallow ReLU network

- Consider **two-layer ReLU network**  $F: \underset{\text{parameter}}{\mathbb{R}^{d_1 \times d_0}} \times \underset{\text{input}}{\mathbb{R}^{d_0}} \rightarrow \underset{\text{prediction}}{\mathbb{R}}$

$$F(W, x) = v^T \sigma(Wx),$$

where  $\sigma: s \mapsto \max\{0, s\}$  componentwise, and  $v \in \mathbb{R}^{d_1}$ .

- To accommodate a bias, we can add a 1 component to  $x$ .

## Shallow ReLU network

- Consider **two-layer ReLU network**  $F: \underset{\text{parameter}}{\mathbb{R}^{d_1 \times d_0}} \times \underset{\text{input}}{\mathbb{R}^{d_0}} \rightarrow \underset{\text{prediction}}{\mathbb{R}}$

$$F(W, x) = v^T \sigma(Wx),$$

where  $\sigma: s \mapsto \max\{0, s\}$  componentwise, and  $v \in \mathbb{R}^{d_1}$ .

- To accommodate a bias, we can add a 1 component to  $x$ .
- This map is piecewise polynomial in  $W, v$  and piecewise linear in  $x$ .

## Activation regions and Jacobian

- For data  $X$ , the smooth pieces are separated by  $\langle w^{(i)}, x^{(j)} \rangle = 0$ .
- For each  $A = [a^{(1)}, \dots, a^{(n)}] \in \{0, 1\}^{d_1 \times n}$  define **activation region**

$$\mathcal{S}_X^A := \left\{ W \in \mathbb{R}^{d_1 \times d_0} : (2A_{ij} - 1) \langle w^{(i)}, x^{(j)} \rangle > 0 \ \forall i \in [d_1], j \in [n] \right\}.$$

Parameters so that  $i$ th unit is active on  $j$ th data point iff  $A_{ij} = 1$ .

## Activation regions and Jacobian

- For data  $X$ , the smooth pieces are separated by  $\langle w^{(i)}, x^{(j)} \rangle = 0$ .
- For each  $A = [a^{(1)}, \dots, a^{(n)}] \in \{0, 1\}^{d_1 \times n}$  define **activation region**

$$\mathcal{S}_X^A := \left\{ W \in \mathbb{R}^{d_1 \times d_0} : (2A_{ij} - 1) \langle w^{(i)}, x^{(j)} \rangle > 0 \ \forall i \in [d_1], j \in [n] \right\}.$$

Parameters so that  $i$ th unit is active on  $j$ th data point iff  $A_{ij} = 1$ .

- The **Jacobian** of the vector of predictions is

$$\nabla_{\theta} F(W, X) = [(v \odot a^{(j)}) \otimes x^{(j)}]_j, \quad \forall W \in \mathcal{S}_X^A, \quad \forall A.$$

- For a fixed input data point  $x^{(i)}$ , a single ReLU

$$w \mapsto \sigma(\langle w, x^{(i)} \rangle)$$

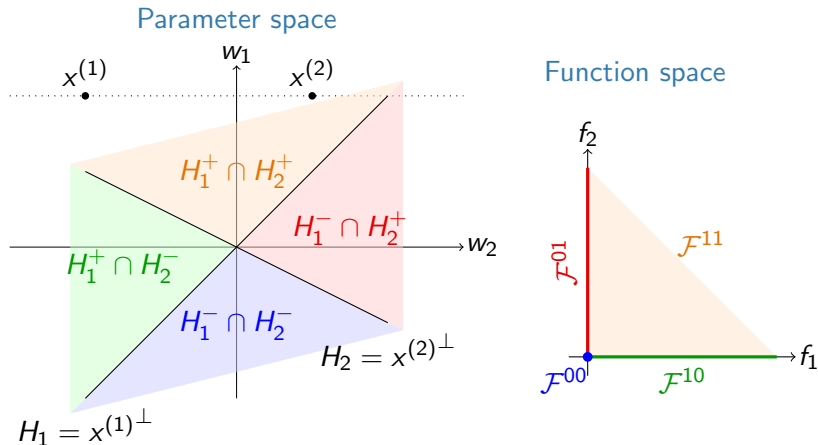
has two activation regions separated by hyperplane  $H_i = x^{(i)\perp}$ .

- This is analog to the linear regions of

$$x \mapsto \sigma(\langle w, x \rangle)$$

in input space for fixed parameter  $w$ .

# Subdivision of parameter space



**Figure 1:** Fan of activation regions; activation patterns indicate the input data points on which each unit is active.

Activation regions with no bad local minima

## Theorem 2 (Most activation regions are good)

Let  $\epsilon > 0$ . If

$$d_1 \geq \max \left( \frac{n}{d_0}, \Omega \left( \log \left( \frac{n}{\epsilon d_0} \right) \right) \right),$$

then for generic datasets  $(X, y)$ , the following holds.

*In all but at most an  $\epsilon$  fraction of all activation regions (i.e. at most  $\lceil \epsilon 2^{d_1} \rceil$ ), every differentiable critical point of  $L$  is a global minimum.*

## Theorem 2 (Most activation regions are good)

Let  $\epsilon > 0$ . If

$$d_1 \geq \max \left( \frac{n}{d_0}, \Omega \left( \log \left( \frac{n}{\epsilon d_0} \right) \right) \right),$$

then for generic datasets  $(X, y)$ , the following holds.

*In all but at most an  $\epsilon$  fraction of all activation regions (i.e. at most  $\lceil \epsilon 2^{d_1} \rceil$ ), every differentiable critical point of  $L$  is a global minimum.*

Caveat: This refers to all activation regions, empty or non-empty.  
More on this later.

## Proof sketch

Theorem 3 (Most binary matrices are full rank, [BVW10])

*Let  $A$  be a  $d \times d$  matrix whose entries are iid random variables sampled uniformly from  $\{0, 1\}$ . Then  $A$  is singular with probability at most*

$$\left( \frac{1}{\sqrt{2}} + o(1) \right)^d.$$

## Proof sketch

### Theorem 3 (Most binary matrices are full rank, [BVW10])

*Let  $A$  be a  $d \times d$  matrix whose entries are iid random variables sampled uniformly from  $\{0, 1\}$ . Then  $A$  is singular with probability at most*

$$\left(\frac{1}{\sqrt{2}} + o(1)\right)^d.$$

### Lemma 4

*Let  $a^{(j)} \in \mathbb{R}^{d_1}$ ,  $x^{(j)} \in \mathbb{R}^{d_0}$  for  $j \in [n]$  and  $v \in \mathbb{R}^{d_1}$ , with  $v_i \neq 0$ ,  $i \in [d_1]$ . Then*

$$\text{rank}(\{(v \odot a^{(j)}) \otimes x^{(j)} : j \in [n]\}) = \text{rank}(\{a^{(j)} \otimes x^{(j)} : j \in [n]\}).$$

Our Jacobian is  $\nabla_{\theta} F(W, X) = [(v \odot a^{(j)}) \otimes x^{(j)}]_j$ ,  $\forall W \in \mathcal{S}_X^A$ ,  $\forall A$ .

## Proof sketch for Theorem 2

- Partition  $[n]$  into  $r \leq d_0$  blocks  $S_1, \dots, S_r$ ,  $|S_k| \leq d_1$ .
- By Theorem 3, each corresponding block of  $A$  fails to have full rank with probability at most  $(\frac{1}{\sqrt{2}} + o(1))^{d_1} \leq C_1 \cdot 0.72^{d_1}$ .
- Using a union bound,

$$\Pr((a^{(s)})_{s \in S_k} \text{ linearly independent for all } k \in [r]) \geq 1 - rC_1 \cdot 0.72^{d_1}.$$

- Then  $A * X$  is full rank for  $X$  with  $x^{(i)} = e_k$  for  $i \in S_k$ . Thus, for most  $A$  the set  $\mathcal{J}^A = \{X \in \mathbb{R}^{d_0 \times n} : A * X \text{ full rank}\}$  is non-empty and, being complement of a Zariski closed set, contain almost every  $X$ .
- If we take  $X \in \mathcal{J} = \cap_{A: \mathcal{J}^A \neq \emptyset} \mathcal{J}^A$  and  $d_1 \geq \log(C_1(n+1)/d_0\epsilon)/\log(1/0.72)$ , then for at least a  $1 - \epsilon$  fraction of all  $A$  we have full rank Jacobian on  $\mathcal{S}_X^A$ . □

## Non-empty activation regions

## Subdivision of parameter space

### Proposition 5 (Number of non-empty regions)

*Consider a network with one layer of  $d_1$  ReLUs. If the columns of  $X$  are in general position in a  $d$ -dimensional linear space, then the number of non-empty activation regions in the parameter space is  $(2 \sum_{k=0}^{d-1} \binom{n-1}{k})^{d_1}$ .*

Regions of a product central hyperplane arrangement.

## Subdivision of parameter space

### Proposition 5 (Number of non-empty regions)

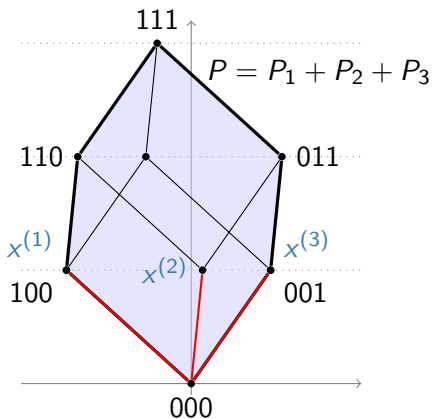
*Consider a network with one layer of  $d_1$  ReLUs. If the columns of  $X$  are in general position in a  $d$ -dimensional linear space, then the number of non-empty activation regions in the parameter space is  $(2 \sum_{k=0}^{d-1} \binom{n-1}{k})^{d_1}$ .*

Regions of a product central hyperplane arrangement.

### Proposition 6 (Identity of non-empty regions)

*Let  $A \in \{0, 1\}^{d_1 \times n}$ . The corresponding activation region is non-empty if and only if  $\sum_{j: A_{ij}=1} x^{(j)}$  is a vertex of  $\sum_{j \in [n]} \text{conv}\{0, x^{(j)}\}$  for all  $i \in [d_1]$ .*

Combination of covectors of the oriented matroid of the input data.



**Figure 2:** The polytope  $P$  of a ReLU on data points  $x^{(1)}, x^{(2)}, x^{(3)}$  is the Minkowski sum of the line segments  $P_i = \text{conv}\{0, x^{(i)}\}$ . The activation regions are the normal cones of  $P$ . The vertices of  $P$  correspond to the non-empty activation regions.

## High-dimensional input

For high-dimensional inputs, most activation regions are non-empty, thus:

### Corollary 7 (Most non-empty activation regions are good)

*Under the same assumptions as Theorem 2, if  $d \geq n$ , then for  $X$  in general position and arbitrary  $y$ :*

*In all but at most an  $\epsilon$  fraction of all **non-empty** activation regions, every differentiable critical point of  $L$  is a zero loss global minimum.*

Non-empty activation regions with no bad local minima

For 1D input, we can explicitly list the non-empty activation regions.

### Lemma 8 (Non-empty activation regions for 1D data)

*Fix a dataset  $(X, y)$  with  $x^{(1)} < x^{(2)} < \dots < x^{(n)}$ . Let  $A \in \{0, 1\}^{d_1 \times n}$ . Then  $S_X^A$  is non-empty if and only if the rows of  $A$  are step vectors. In particular, there are exactly  $(2n)^{d_1}$  non-empty activation regions.*

### Theorem 9 (Most non-empty activation regions are good)

*Let  $\epsilon \in (0, 1)$ . Suppose that  $X$  consists of distinct data points, and*

$$d_1 \geq 2n \log \left( \frac{n}{\epsilon} \right).$$

*Then in all but at most an  $\epsilon$  fraction of non-empty activation regions,  $\nabla_{\theta} F$  is full rank and every differentiable critical point of  $L$  is a global minimum.*

## Proof sketch

### Lemma 10 (Coupon collector's problem)

*Let  $\epsilon \in (0, 1)$ , and let  $n \leq m$  be positive integers. Let  $C_1, C_2, \dots, C_d \in [m]$  be iid random variables such that for all  $i \in [d]$  one has  $\mathbb{P}(C_i = j) \geq \delta$ . If*

$$d \geq \frac{1}{\delta} \log \left( \frac{n}{\epsilon} \right),$$

*then  $[n] \subseteq \{C_1, \dots, C_d\}$  with probability at least  $1 - \epsilon$ .*

## Theorem 11 (Fraction of regions with global minima)

Let  $\epsilon \in (0, 1)$ . Suppose that  $X$  consists of distinct data points, and

$$|\{i \in [d_1] : v^{(i)} > 0\}| \geq 2n \log \left( \frac{2n}{\epsilon} \right),$$

and

$$|\{i \in [d_1] : v^{(i)} < 0\}| \geq 2n \log \left( \frac{2n}{\epsilon} \right).$$

Then in all but at most an  $\epsilon$  fraction of non-empty activation regions  $S_X^A$ , the subset of global minimizers  $\mathcal{G}_{X,Y} \cap S_X^A$  is a non-empty affine set of codimension  $n$ . Moreover, all global minima of  $L$  have zero loss.

## Function space on 1D data

## Proposition 12 (Function space on one-dimensional data)

Let  $X$  be a list of  $n$  distinct points in  $1 \times \mathbb{R}$  with  $x^{(1)} < x^{(2)} < \dots < x^{(n)}$ . Let  $\bar{x}^{(i)} = [x_2^{(i)}, -1]$  and  $X_{\geq i} = [0, \dots, 0, x^{(i)}, \dots, x^{(n)}]$ .

- Then the functions a ReLU represents on  $X$  form a *polyhedral cone*,  $\alpha f \in \mathbb{R}^n$  with  $\alpha \geq 0$  and  $f$  in the polyline with vertices

$$\bar{x}^{(i)} X_{\leq i}, \quad i = 1, \dots, n \quad \text{and} \quad -\bar{x}^{(i)} X_{\geq i}, \quad i = 1, \dots, n. \quad (2)$$

- A sum of  $m$  ReLUs represents non-negative scalar multiples of *convex combinations* of any  $m$  points on this polyline.
- Arbitrary linear combinations of  $m$  ReLUs represent scalar multiples of *affine combinations* of any  $m$  points on this polyline.

## Function space

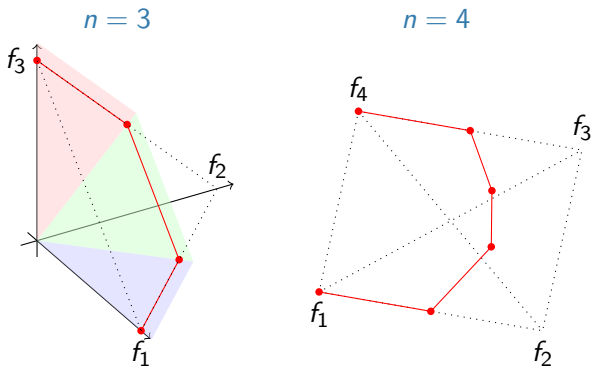


Figure 3: Function space of a ReLU on  $n$  data points in  $1 \times \mathbb{R}$ , for  $n = 3, 4$ .

# Experiments

# Probability of full rank Jacobian for random init

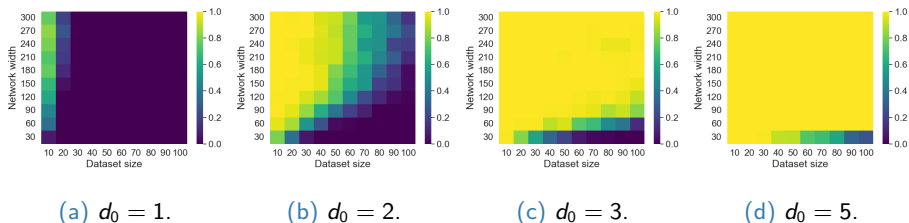
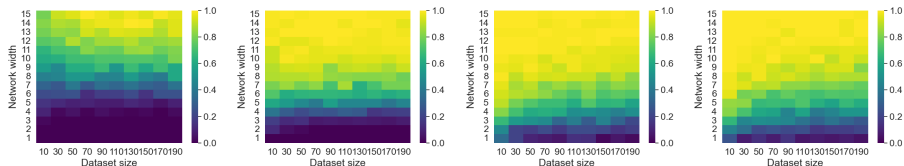


Figure 4: Input dimension  $d_0$  is left fixed. Minimum  $d_1$  to achieve full rank linear in  $n$ , slope decreases as  $d_0$  increases, as predicted by Theorem 2.

# Probability of full rank Jacobian for random init



(a)  $d_0 = \lceil \frac{n}{4} \rceil$ .

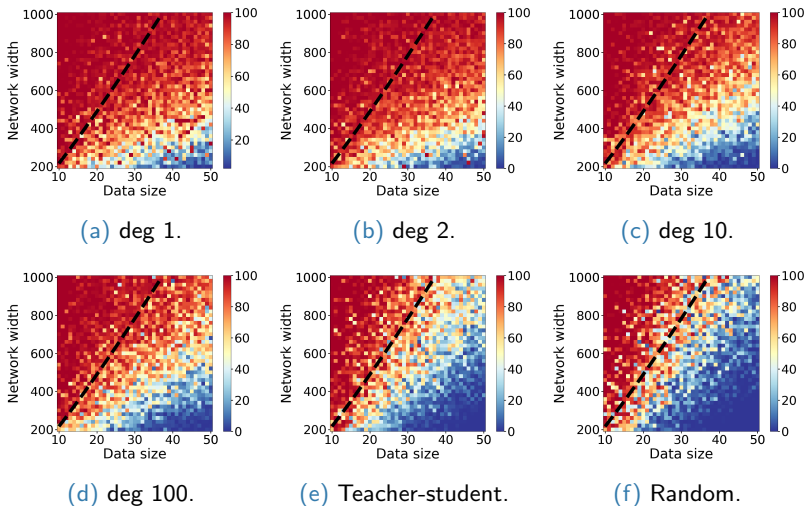
(b)  $d_0 = \lceil \frac{n}{2} \rceil$ .

(c)  $d_0 = n$ .

(d)  $d_0 = 2n$ .

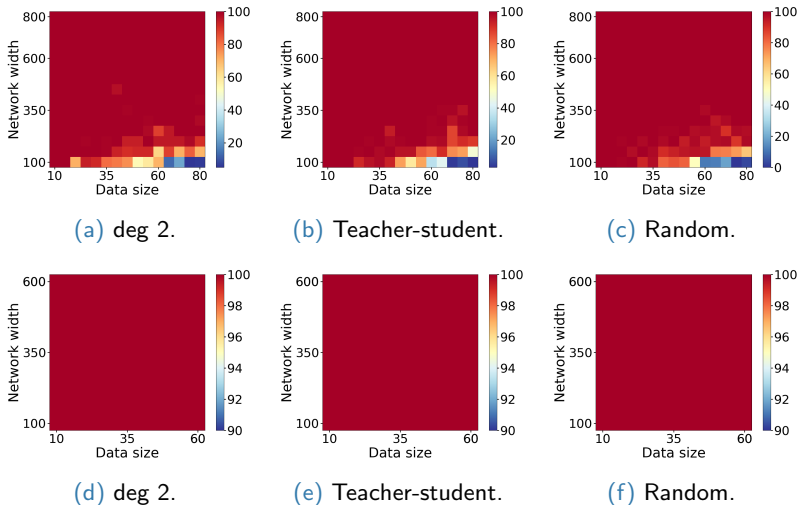
Figure 4: Input dimension  $d_0$  scales linearly in the number of samples  $n$ . Minimum  $d_1$  to achieve full rank remains constant in  $n$ , consistent with Theorem 2.

## Percentage of regions with global min, $d_0 = 1$



**Figure 5:** Percentage of randomly sampled activation regions that contain a global minimum of the loss for networks with  $d_0 = 1$ . Black line is Theorem 11.

## Percentage of regions with global min, $d_0 = 2, 5$



**Figure 6:** Percentage of randomly sampled activation regions that contain a global minimum for networks with input dimension  $d_0 = 2$  (top) and  $d_0 = 5$  (bottom). Consistent with Theorem 2 and Corollary 7.

## Summary

- We studied the loss landscape of two-layer ReLU networks in the mildly overparameterized regime.
- Most activation regions have no bad differentiable local minima.
- In the univariate case, most non-empty activation regions contain a high-dimensional set of global minimizers.

## Further topics

- Gradient descent.
- Non-empty regions for multivariate data.
- Deep networks.

## Questions

- Properties of oriented matroids (as matrices) of given datasets.
- Volume of normal cones and parameter initialization.
- Function spaces of deep networks.

# References I



Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang.

On exact computation with an infinitely wide neural net.

In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.



Peter Auer, Mark Herbster, and Manfred K. K Warmuth.

Exponentially many local minima for single neurons.

In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995.



Jean Bourgain, Van H Vu, and Philip Matchett Wood.

On the singularity probability of discrete random matrices.

*Journal of Functional Analysis*, 258(2):559–603, 2010.



Kenji Fukumizu and Shun-ichi Amari.

Local minima and plateaus in hierarchical structures of multilayer perceptrons.

*Neural Networks*, 13(3):317–327, 2000.

## References II



Arthur Jacot, Franck Gabriel, and Clement Hongler.

Neural tangent kernel: Convergence and generalization in neural networks.  
*In Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.



Kedar Karhadkar, Michael Murray, Hanna Tseran, and Guido Montúfar.

Mildly overparameterized ReLU networks have a favorable loss landscape.  
*arXiv:2305.19510*, 2023.



Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri.

Deep neural networks as Gaussian processes.  
*In International Conference on Learning Representations*, 2018.



Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington.

Wide neural networks of any depth evolve as linear models under gradient descent.  
*In Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

## References III



Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar.  
Characterizing the spectrum of the NTK via a power series expansion.  
*In The Eleventh International Conference on Learning Representations, 2023.*



Radford M. Neal.  
*Bayesian Learning for Neural Networks.*  
Springer-Verlag, Berlin, Heidelberg, 1996.



Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro.  
In search of the real inductive bias: On the role of implicit regularization in deep learning.  
*In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings, 2015.*



Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu.  
Local minima in training of deep networks, 2017.

## References IV



Eduardo Sontag and Héctor J. Sussmann.

Backpropagation can give rise to spurious local minima even for networks without hidden layers.

*Complex Syst.*, 3, 1989.



Itay Safran and Ohad Shamir.

Spurious local minima are common in two-layer ReLU neural networks.

In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4433–4441. PMLR, 10–15 Jul 2018.



Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro.

Kernel and rich regimes in overparametrized models.

In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 2020.