Deep Learning - Parameters and Functions

Mildly Overparametrized ReLU Nets

Guido Montúfar
montufar@math.ucla.edu

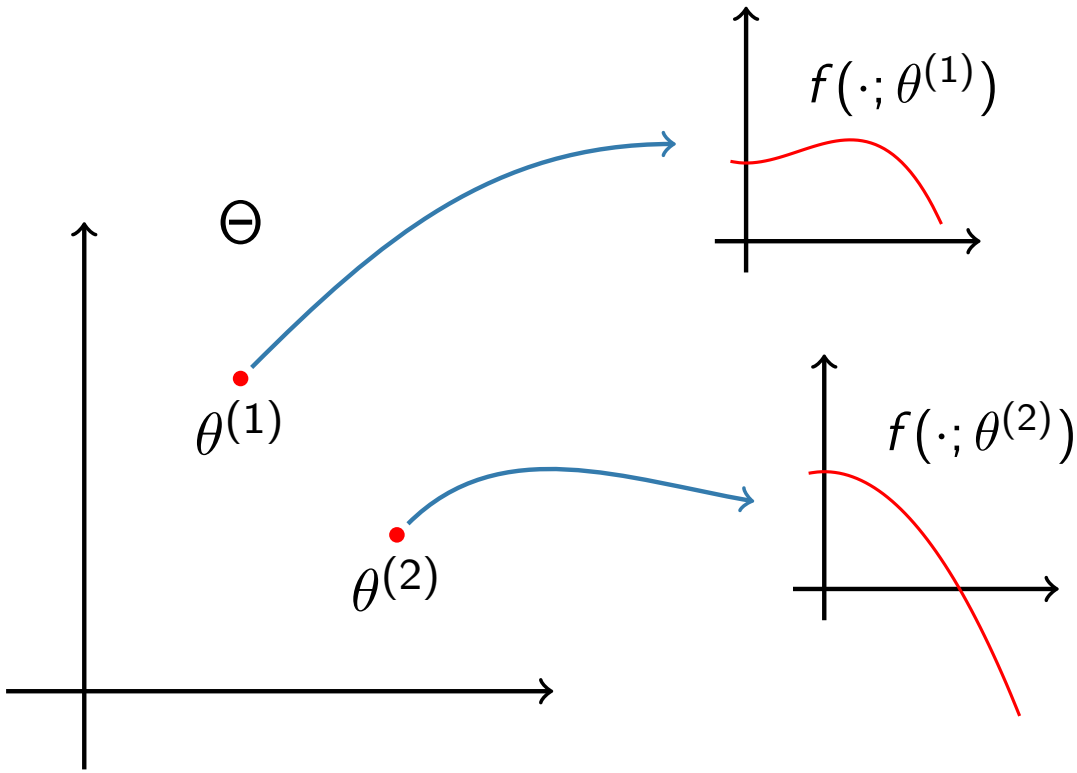
48th Winter Conference in Statistics, March 2024, Hemavan

Kedar Karhadkar

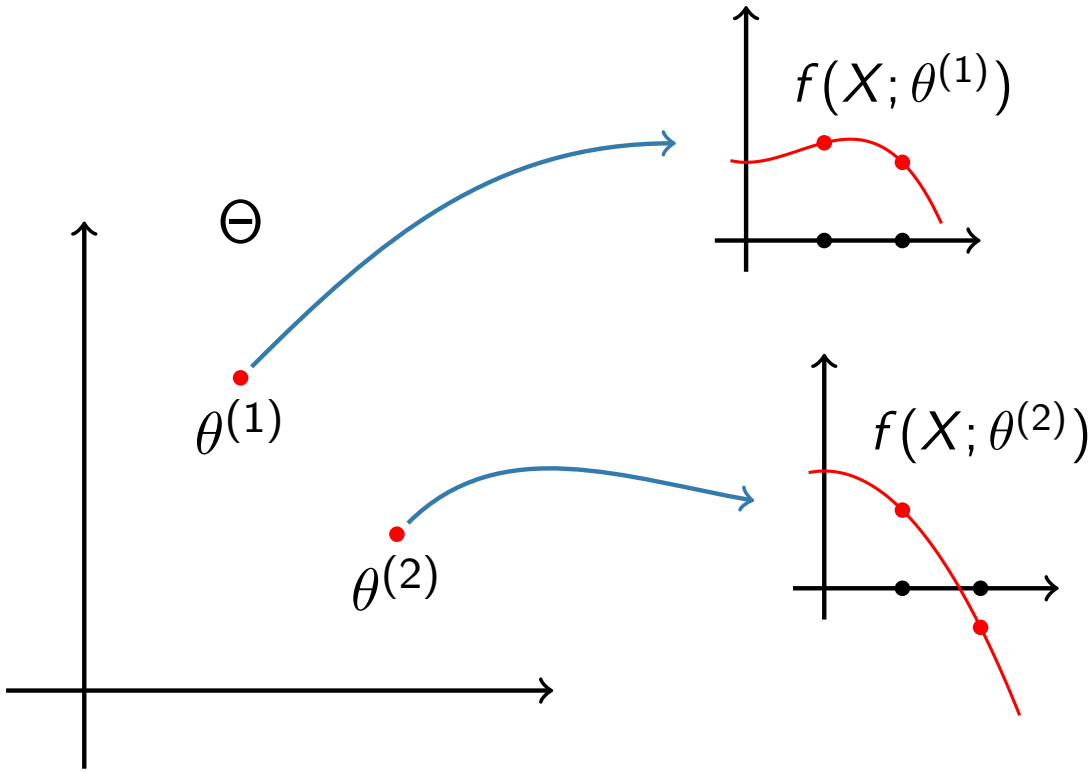
Michael Murray

Hanna Tseran

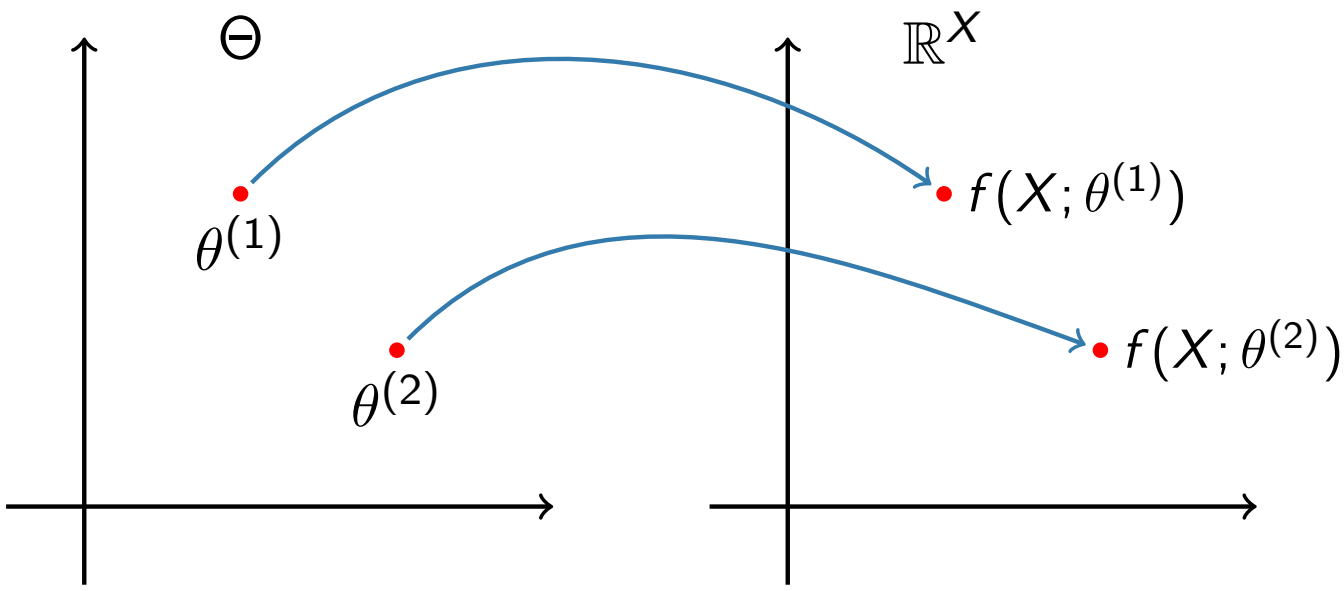
- "Mildly Overparameterized ReLU Networks Have a Favorable Loss Landscape"



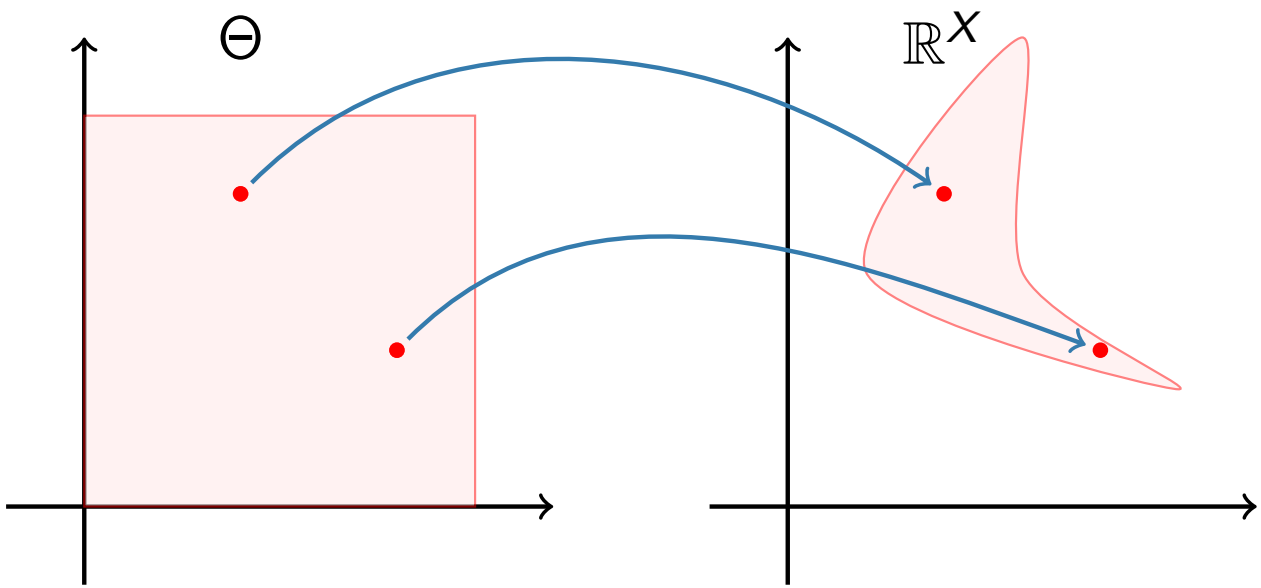
Parametric model



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[1].

[1] Auer, Herbster, and Warmuth 1995; Fukumizu and Amari 2000; Safran and Shamir 2018; Sontag and Sussmann 1989; Swirszcz, Czarnecki, and Pascanu 2017.

Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus[1].
- A puzzling question is why bad local minima do not seem to be a problem for training.

[1] Auer, Herbster, and Warmuth 1995; Fukumizu and Amari 2000; Safran and Shamir 2018; Sontag and Sussmann 1989; Swirszcz, Czarnecki, and Pascanu 2017.

Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus[1].
- A 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.

[1] Auer, Herbster, and Warmuth 1995; Fukumizu and Amari 2000; Safran and Shamir 2018; Sontag and Sussmann 1989; Swirszcz, Czarnecki, and Pascanu 2017.

Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus[1].
- A 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 can 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.

[1] Auer, Herbster, and Warmuth 1995; Fukumizu and Amari 2000; Safran and Shamir 2018; Sontag and Sussmann 1989; Swirszcz, Czarnecki, and Pascanu 2017.

Overview

- Neural networks have a non-convex loss landscape with local minima and plateaus[1].
- A 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 can 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.

[1] Auer, Herbster, and Warmuth 1995; Fukumizu and Amari 2000; Safran and Shamir 2018; Sontag and Sussmann 1989; Swirszcz, Czarnecki, and Pascanu 2017.

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 ϵ fraction, have no bad local minima.
- For generic high-dimensional input data $d_0 \geq n$, most *non-empty* activation regions will have no bad local minima. Extension to deep case.

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 ϵ fraction, have no bad local minima.
- For generic high-dimensional input data $d_0 \geq n$, most *non-empty* activation regions will have no bad local minima. Extension to deep case.
- Theorem 7: If $d_0 = 1$ and $d_1 = \Omega(n \log(\frac{n}{\epsilon}))$, all but at most an ϵ 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 ϵ fraction, have no bad local minima.
- For generic high-dimensional input data $d_0 \geq n$, most *non-empty* activation regions will have no bad local minima. Extension to deep case.
- Theorem 7: If $d_0 = 1$ and $d_1 = \Omega(n \log(\frac{n}{\epsilon}))$, all but at most an ϵ fraction of *non-empty* activation regions have no bad local minima.
- Theorem 8: If $d_0 = 1$ and $d_1 = d_+ + d_-$ with $d_+, d_- = \Omega(n \log(\frac{n}{\epsilon}))$, then all but at most an ϵ fraction of non-empty activation regions contain an affine set of global minima of codimension n.

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 ϵ fraction, have no bad local minima.
- For generic high-dimensional input data $d_0 \geq n$, most *non-empty* activation regions will have no bad local minima. Extension to deep case.
- Theorem 7: If $d_0 = 1$ and $d_1 = \Omega(n \log(\frac{n}{\epsilon}))$, all but at most an ϵ fraction of *non-empty* activation regions have no bad local minima.
- Theorem 8: If $d_0 = 1$ and $d_1 = d_+ + d_-$ with $d_+, d_- = \Omega(n \log(\frac{n}{\epsilon}))$, then all but at most an ϵ fraction of non-empty activation regions contain an affine set of global minima of codimension n.
- Theorem 12 provide bounds on the fraction of regions with bad local minima by volume.

- We consider input and output data

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

- 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

$$F\colon \underset{\textit{parameter}}{\mathbb{R}^m} \times \underset{\textit{input}}{\mathbb{R}^d} \to \underset{\textit{prediction}}{\mathbb{R}}$$

and the vector of predictions on input data X,

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

- 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

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

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 $L\colon \underset{\text{parameter}}{\mathbb{R}^m} \times \underset{\text{inputs}}{\mathbb{R}^{d \times n}} \times \underset{\text{outputs}}{\mathbb{R}^{1 \times n}} \to \mathbb{R}^1$,

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

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

Proof.

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

□

ReLU network

- E.g., a two-layer ReLU network $F\colon \underset{\textit{parameter}}{\mathbb{R}^{d_1\times d_0}} \times \underset{\textit{input}}{\mathbb{R}^{d_0}} \to \underset{\textit{prediction}}{\mathbb{R}}$

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

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

ReLU network

- E.g., a two-layer ReLU network $F\colon \underset{\textit{parameter}}{\mathbb{R}^{d_1\times d_0}} \times \underset{\textit{input}}{\mathbb{R}^{d_0}} \to \underset{\textit{prediction}}{\mathbb{R}}$

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

  where $\sigma\colon s\mapsto \max\{0,s\}$ componentwise, and $v\in\mathbb{R}^{d_1}$.
- To accommodate a bias, we can add a 1 component to x.

ReLU network

- E.g., a two-layer ReLU network $F\colon \underset{\text{parameter}}{\mathbb{R}^{d_1\times d_0}} \times \underset{\text{input}}{\mathbb{R}^{d_0}} \to \underset{\text{prediction}}{\mathbb{R}}$

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

  where $\sigma\colon s\mapsto \max\{0,s\}$ 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 $A = [a^{(1)}, \ldots, a^{(n)}] \in \{0,1\}^{d_1 \times n}$ 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 ith unit is active on jth 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 $A = [a^{(1)}, \ldots, a^{(n)}] \in \{0,1\}^{d_1 \times n}$ 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 ith unit is active on jth data point iff $A_{ij} = 1$.
- The Jacobian of the vector of predictions is

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

- Similar definitions for deep ReLU nets.

Subdivision of parameter space

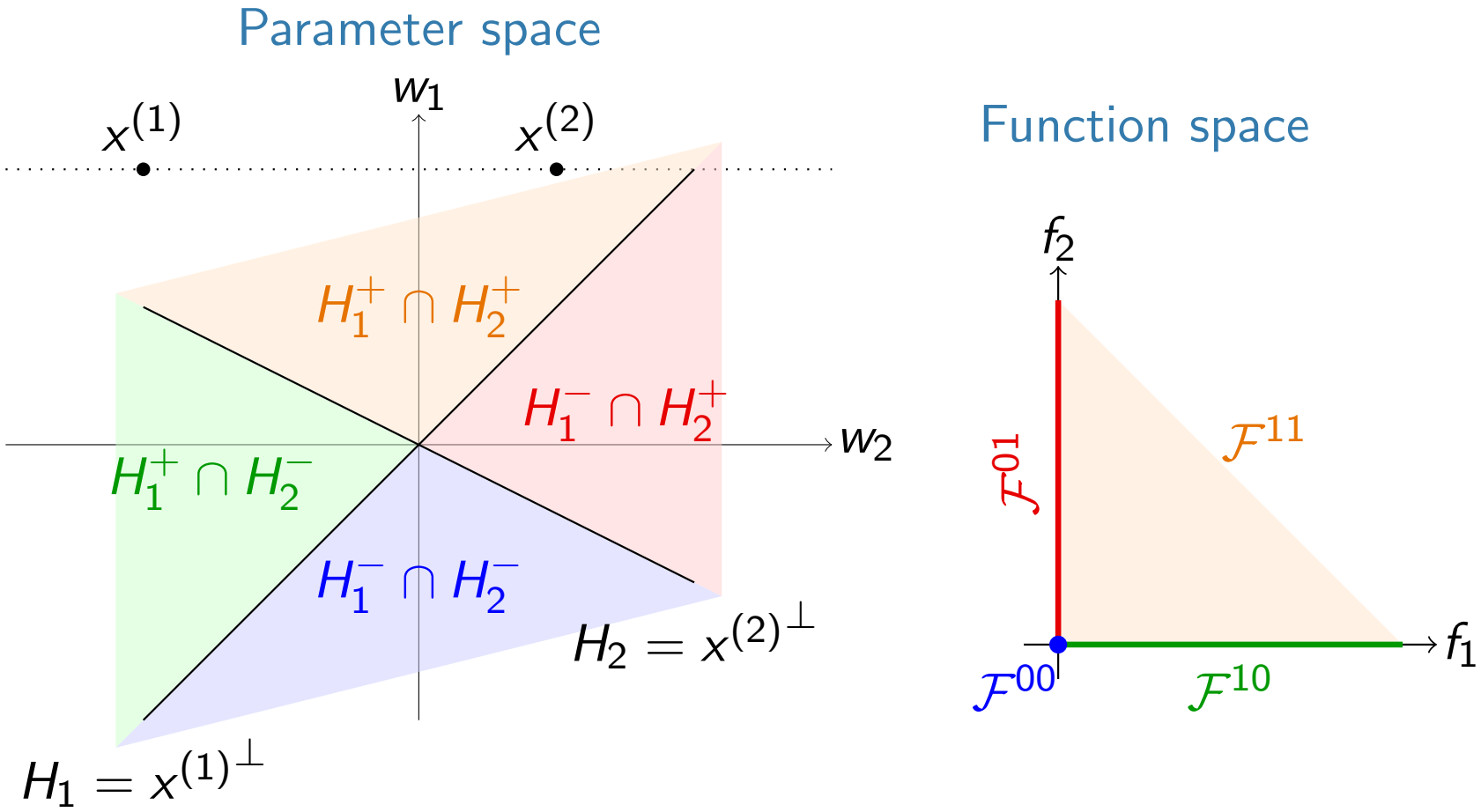


Figure 1: Fan of activation regions; activation patterns indicate the input data points on which each unit is active.

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 ϵ *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 ϵ *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.*

Uses an upper bound on probability that a binary random matrix is singular.
Caveat: This refers to all activation regions, empty or non-empty.

Non-empty activation regions

Proposition 3 (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 3 (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 4 (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 $\sum_{j\in[n]} \mathrm{conv}\{0, x^{(j)}\}$ for all $i \in [d_1]$.

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

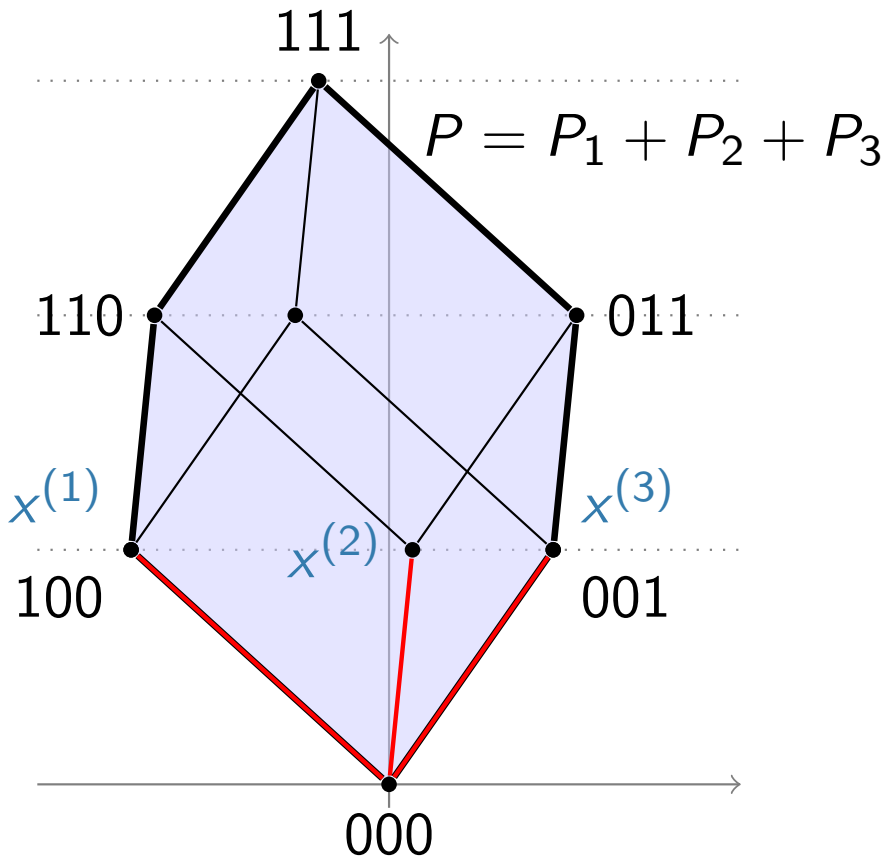


Figure 2: The polytope P of a ReLU on data points $x^{(1)}, x^{(2)}, x^{(3)}$ is the Minkowski sum of the line segments $P_i = \text{conv}\{0, x^{(i)}\}$.
The activation regions are the normal cones of P.
The vertices of P correspond to the non-empty activation regions.

High-dimensional input

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

Corollary 5 (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 ϵ 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 6 (Non-empty activation regions for 1D data)

Fix a dataset (X, y) *with* $x^{(1)} < x^{(2)} < \cdots < x^{(n)}$. *Let* $A \in \{0,1\}^{d_1 \times n}$. *Then* $\mathcal{S}_X^A$ *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 7 (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 ϵ fraction of non-empty activation regions, $\nabla_\theta F$ is full rank and every differentiable critical point of L is a global minimum.

Theorem 8 (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 ϵ fraction of non-empty activation regions $\mathcal{S}_X^A$, the subset of global minimizers $\mathcal{G}_{X,y} \cap \mathcal{S}_X^A$ is a non-empty affine set of codimension n. Moreover, all global minima of L have zero loss.

We can extend the analysis to handle points on the boundaries between regions, where the loss is non-differentiable.

Theorem 9

Let $\epsilon \in (0,1)$. *If*

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

then in all but at most a fraction ϵ *of non-empty activation regions* A, *every local minimum of* L *in* $\mathcal{S}_X^A \times \mathbb{R}^{d_1}$ *is a global minimum.*

Extension to deep nets

- For $l \in \{0, \dots, L\}$, we define the l-th layer $f_l : \mathbb{R}^m \times \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ by

$$\begin{aligned} f_0(W, x) :=& x, \\ f_l(W, x) :=& \sigma(W_l f_{l-1}(\theta, x)) \quad \text{if } l \in [L-1], \text{ and} \\ f_L(W, x) :=& v^T f_{l-1}(\theta, x), \end{aligned}$$

  where $v \in \mathbb{R}^{d_{l-1}}$ is a fixed vector whose entries are nonzero.
- The activation patterns of a deep network are given by tuples

$$A = (A_1, A_2, \dots, A_{L-1}),$$

  where for each $l \in [L-1]$, $A_l \in \{0,1\}^{d_l \times n}$.
- Denote $\mathcal{S}_X^A$ subset of parameters with activation pattern A.

Theorem 10

Let $X \in \mathbb{R}^{d_0 \times n}$ be an input dataset with distinct points. Suppose that for all $l \in [L-2]$,

$$d_l = \Omega\left(\log \frac{n}{\epsilon L}\right),$$

and that

$$d_{L-1} = n + \Omega\left(\log \frac{1}{\epsilon}\right).$$

Then for at least a $(1-\epsilon)$ fraction of all activation patterns A, the following holds. For all $W \in \mathcal{S}_X^A$, $\nabla_W F(W, X)$ has rank n.

Volumes of activation regions

One-dimensional input data

We bound the volume of activation regions with full rank Jacobian in terms of the amount of separation between the data points.

Proposition 11

Let $n \geq 2$. Suppose the entries of v are nonzero. Suppose that for all $j, k \in [n]$ with $j \neq k$, we have $|x^{(j)}| \leq 1$ and $|x^{(j)} - x^{(k)}| \geq \phi$. If

$$d_1 \geq \frac{4}{\phi} \log\left(\frac{n}{\epsilon}\right),$$

then, writing μ for the Lebesgue measure,

$$\mu(\cup\{\mathcal{S}_X^A \cap [-1,1]^{d_1} \times [-1,1]^{d_1} : \nabla_{w,b} F \text{ has full rank on } \mathcal{S}_X^A\}) \geq (1-\epsilon)2^{2d_1}.$$

Arbitrary dimension input data

We say that an input dataset $X \in \mathbb{R}^{d_0 \times n}$ is γ-anticoncentrated if for all nonzero $u \in \mathbb{R}^n$, $\mathbb{P}_{a \sim \mathcal{D}_X}(u^T a = 0) \leq 1 - \gamma$. We can interpret this as a condition on the amount of separation between data points.

Theorem 12

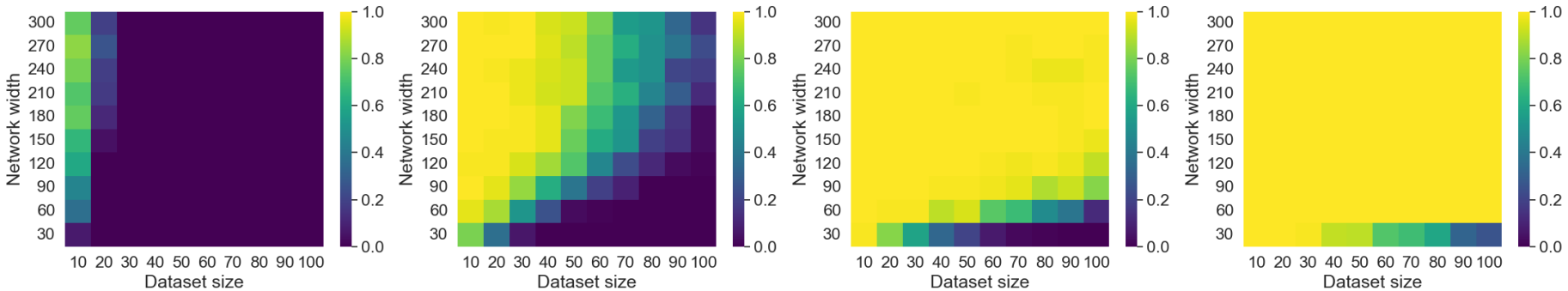
Let $\epsilon, \gamma \in (0,1)$. Suppose that $X \in \mathbb{R}^{d_0 \times n}$ is generic and γ-anticoncentrated. If

$$d_1 \geq \frac{8}{\gamma^2} \log\left(\frac{d_0}{\epsilon}\right) + \frac{2}{\gamma}\left(\frac{n}{d_0} + 1\right),$$

then with probability at least $1 - \epsilon$, $\nabla_{(W,v)} F(W, v, X)$ has rank n.

Experiments

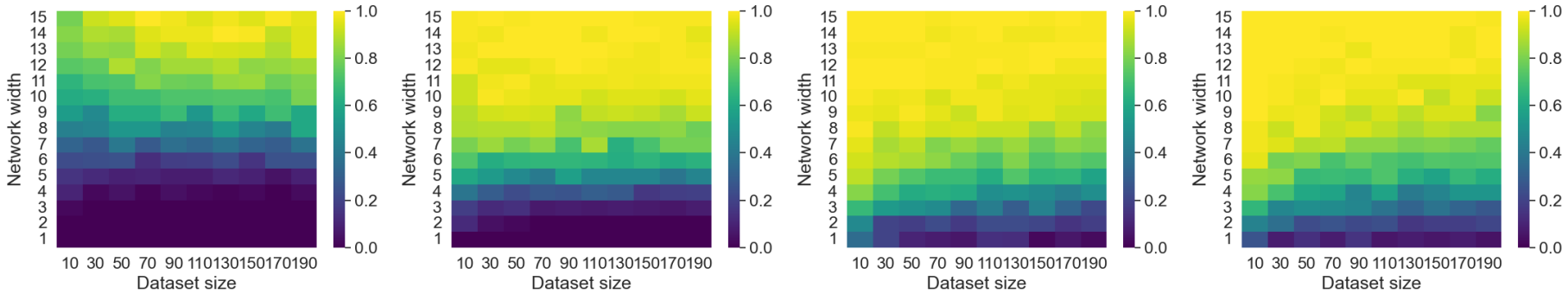
Probability of full rank Jacobian for random init



$d_0 = 1$ $d_0 = 2$ $d_0 = 3$ $d_0 = 5$

Figure 3: 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



$d_0 = \lceil \frac{n}{4} \rceil$ $d_0 = \lceil \frac{n}{2} \rceil$ $d_0 = n$ $d_0 = 2n$

Figure 3: 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, $d_0 = 1$

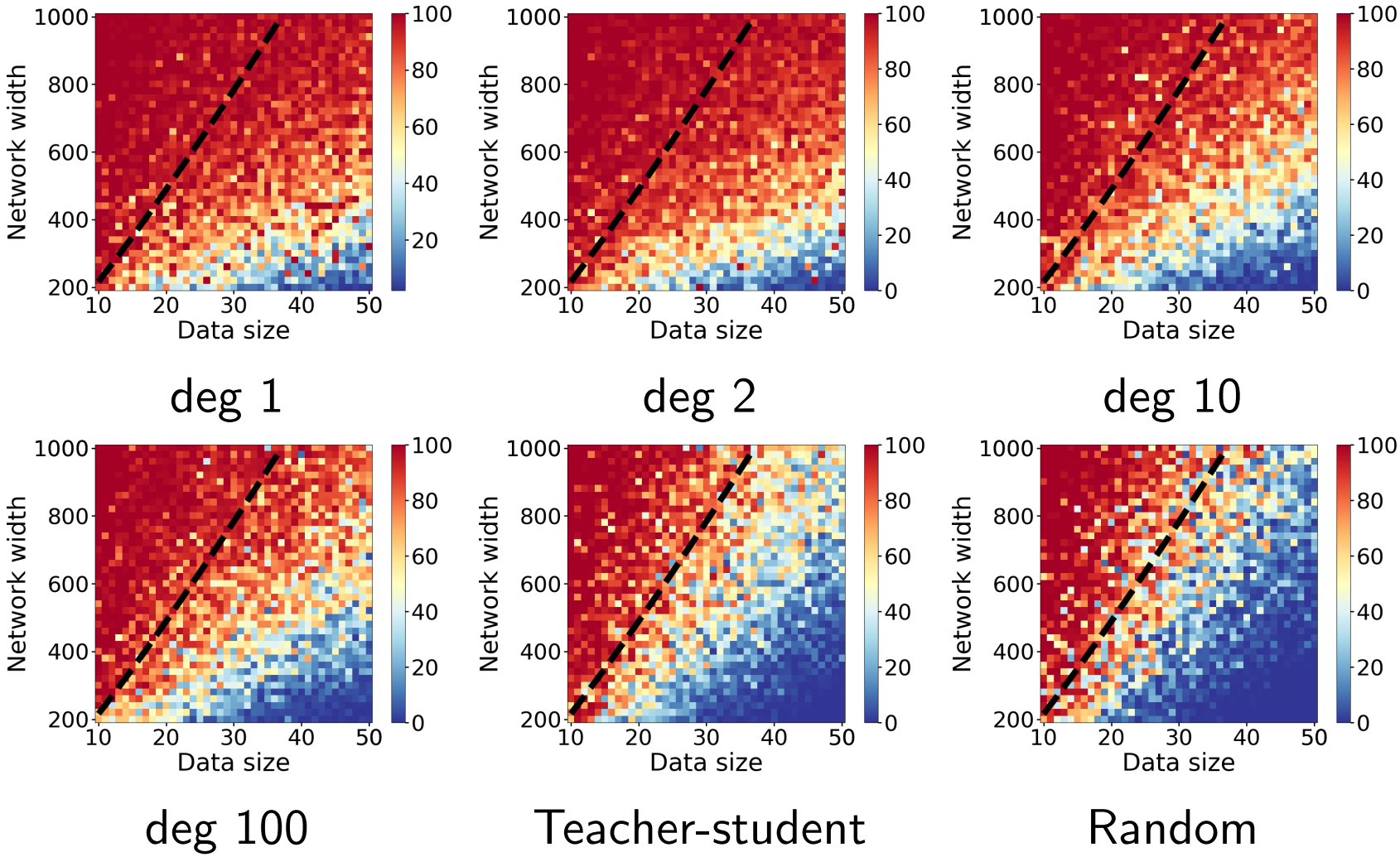


Figure 4: Percentage of randomly sampled activation regions that contain a global minimum of the loss for networks with $d_0 = 1$. Black line is Theorem 8.

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

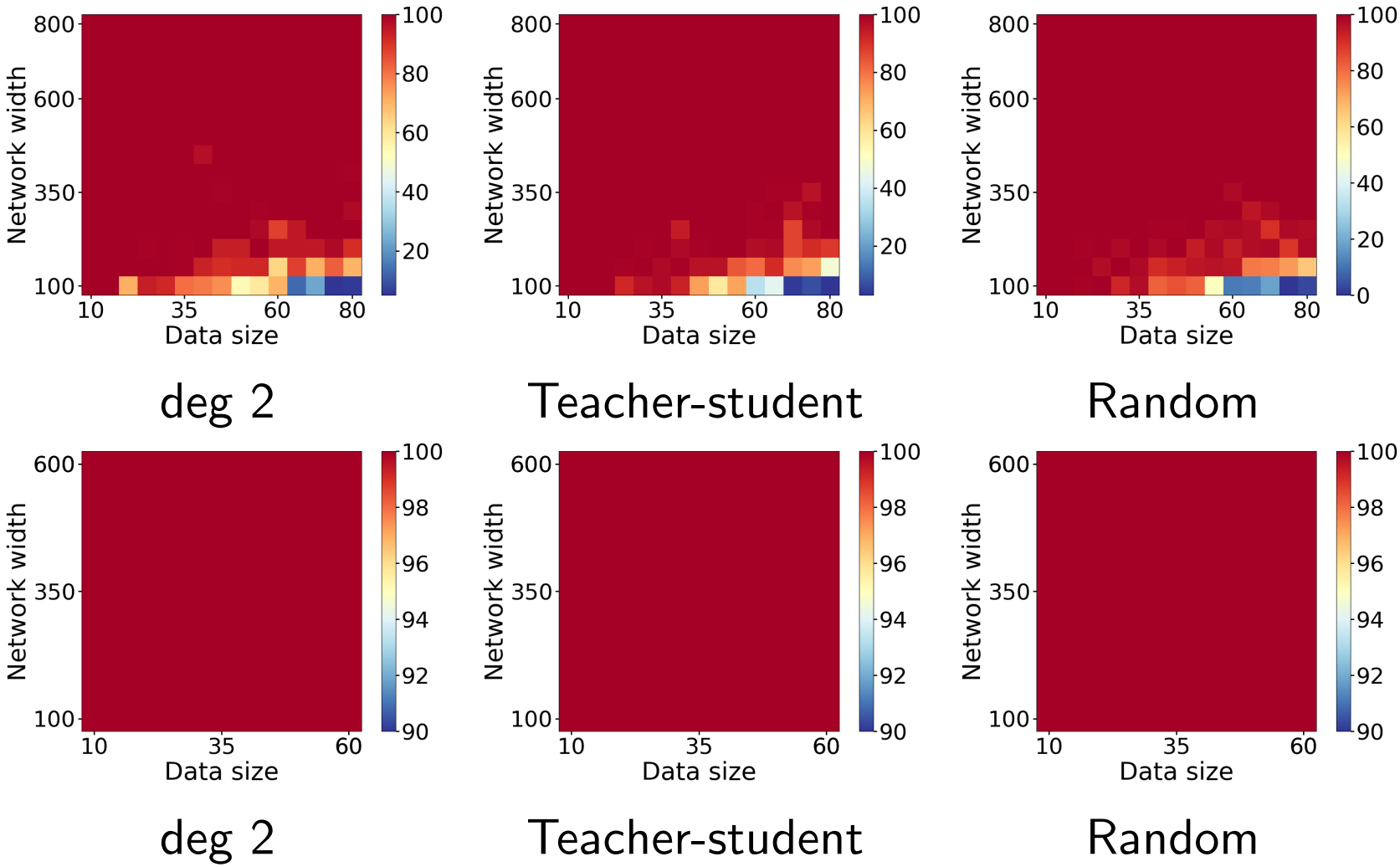


Figure 5: Percentage of randomly sampled activation regions that contain a global minimum for networks with input dimension $d_0 = 2$ (top) and $d_0 = 5$ (bottom). Consistent with Theorem 2 and Corollary 5.

Function space on 1D data

Proposition 13 (Function space on one-dimensional data)

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

- *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*

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

- *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.*

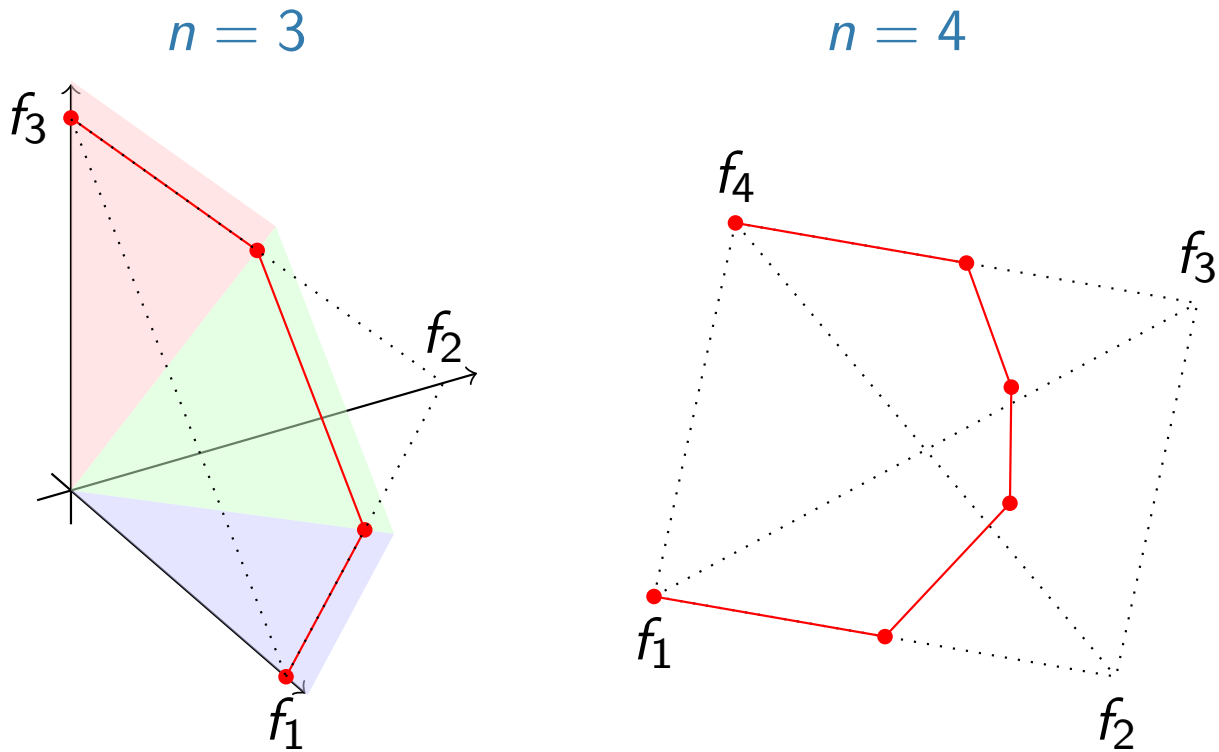


Figure 6: Function space of a ReLU on n data points in $1 \times \mathbb{R}$, for $n = 3, 4$.

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.

References I

Auer, Peter, Mark Herbster, and Manfred K. K Warmuth (1995). "Exponentially many local minima for single neurons". In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky, M.C. Mozer, and M. Hasselmo. Vol. 8. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/1995/file/3806734b256c27e41ec2c6bffa26d9e7-Paper.pdf.

Bourgain, Jean, Van H Vu, and Philip Matchett Wood (2010). "On the singularity probability of discrete random matrices". In: *Journal of Functional Analysis* 258.2, pp. 559–603. URL: https://www.sciencedirect.com/science/article/pii/S0022123609001955.

Fukumizu, Kenji and Shun-ichi Amari (2000). "Local minima and plateaus in hierarchical structures of multilayer perceptrons". In: *Neural Networks* 13.3, pp. 317–327. URL: https://www.sciencedirect.com/science/article/pii/S0893608000000095.

Karhadkar, Kedar et al. (2023). "Mildly Overparameterized ReLU Networks Have a Favorable Loss Landscape". In: *arXiv:2305.19510*.

References II

Safran, Itay and Ohad Shamir (2018). "Spurious Local Minima are Common in Two-Layer ReLU Neural Networks". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 4433–4441. URL: https://proceedings.mlr.press/v80/safran18a.html.

Sontag, Eduardo and Héctor J. Sussmann (1989). "Backpropagation Can Give Rise to Spurious Local Minima Even for Networks without Hidden Layers". In: *Complex Syst.* 3. URL: https://www.complex-systems.com/abstracts/v03_i01_a07/.

Swirszcz, Grzegorz, Wojciech Marian Czarnecki, and Razvan Pascanu (2017). *Local minima in training of deep networks*. URL: https://openreview.net/forum?id=Syoiqwcxx.