Mathematics of Deep Learning
Lecture 6: Over-parametrization, Convergence of GD to Global Minima, Concentration Inequalities, Matrix Perturbation

Prof. Predrag R. Jelenković

Time: Tuesday 4:10-6:40pm

Dept. of Electrical Engineering
Columbia University , NY 10027, USA
Office: 812 Schapiro Research Bldg.
Phone: (212) 854-8174
Email: predrag@ee.columbia.edu
URL: http://www.ee.columbia.edu/∼predrag

High-Dimensional Geometry

Reflection on high-dimensional geometry

- Arguments about high dimensions are often involved, mostly as heuristic, in NNs/ML papers
- But, one has to be careful
  - Counterintuitive: low dimensional intuition does not work
    - Volume of n-cube and n-ball is not found where one would expect it to be
    - Many more vectors are orthogonal than you would think
- Better understanding of the high-dimensional geometry can lead to better statistical learning algorithms.
  - This is exploited in sparse grids, where certain volumes in high dimensions can be ignored.

n-Cube

Define n-cube, $C^n(s)$, with side s, which is centered at the origin

$$C^n(s) = \{(x_1, \ldots, x_n) : -s/2 \le x_i \le s/2, 1 \le i \le n\}$$

Let $C^n = C^n(1)$ be the unit cube.

- The cube C^n has diameter $\sqrt{n}$, but volume 1.
- For any $t > 0$,

$$\lim_{n\to\infty} \mathrm{Vol}(C^n - C^n(1 - t/n)) = \lim_{n\to\infty} (1 - (1 - t/n)^n) = 1 - e^{-t},$$

- Algebraically these are trivial, but geometrically, these is counterintuitive(!)

  Example For $t = 3$ and $n = 300$,

$$1 - (1 - 3/300)^{300} \approx 95\%$$

  of the volume is in a shell of width 0.01

n-Cube: Volume is concentrated in the middle

Consider a hyperplane that goes through the middle (origin)

$$x_1 + \cdots + x_n = 0$$

Note that this hyperplane is perpendicular to the unit vector

$$\frac{1}{\sqrt{n}}(1, \ldots, 1)$$

Let A be the set of points that are at distance c from this hyperplane

$$A = \{(x_1, \ldots, x_n) : \frac{|x_1 + \cdots + x_n|}{\sqrt{n}} \leq c\}$$

Lemma For any $c > 0$, Vol$(C^n \cap A) \geq 1 - e^{-\Omega(c^2)}$

How can this be?

- Most of the volume is both in the middle & thin shell?

n-Ball: Volume is concentrated in the very thin shell

Define an n-Ball and n-Sphere of radius r

$$B^n(r) = \{(x_1, \ldots, x_n) : x_1^2 + \cdots x_n^2 \le r^2\}$$
$$S^{n-1}(r) = \{(x_1, \ldots, x_n) : x_1^2 + \cdots x_n^2 = r^2\}$$

Let $B^n = B^n(1)$ and $S^{n-1} = S^{n-1}(1)$ be the unit ball and unit sphere, respectively.

Lemma Most of the volume is in the thin shell of order t/n,

$$\lim_{n\to\infty} \frac{\mathsf{Vol}(B^n) - \mathsf{Vol}(B^n(1-t/n))}{\mathsf{Vol}(B^n)} = 1 - \lim_{n\to\infty}(1-t/n)^n = 1 - e^{-t}$$

Proof Follows from scaling

$$\mathsf{Vol}(B^n(r)) = r^n \mathsf{Vol}(B^n)$$

(Justify this)

Uniform Random Variables on n-Sphere & n-Ball

Let $X_i, i \geq 1$ be independent standard normal, $\mathcal{N}(0,1)$, random variables and let

$$\|\boldsymbol{X}\|_2 = \sqrt{X_1^2 + \cdots X_n^2}$$

Then,

$$\boldsymbol{Y} = \left(\frac{X_1}{\|\boldsymbol{X}\|_2}, \ldots, \frac{X_n}{\|\boldsymbol{X}\|_2} \right)$$

is uniformly distributed on unit sphere S^{n-1}.

Furthermore, if U is a uniform random variable on $[0,1]$, which is independent of $\boldsymbol{X}$, then

$$\boldsymbol{Z} = \left(\frac{U^{1/n} X_1}{\|\boldsymbol{X}\|_2}, \ldots, \frac{U^{1/n} X_n}{\|\boldsymbol{X}\|_2} \right)$$

is uniformly distributed inside the unit ball B^n.

The preceding representations can be used to prove a variety of interesting results.

n-Ball: Volume is in the middle - around the equator

Lemma For any $c > 0$,

$$\frac{\mathsf{Vol}(B^n \cap \{|x_1| \le c/\sqrt{n}\})}{\mathsf{Vol}(B^n)} \ge 1 - e^{-\Omega(c^2)}$$

Proof We will use the preceding construction of uniform random variables on a sphere. Note first that, with very high probability, $1 - O(e^{-\Omega(n)})$, we can chose a constant γ such that

$$\|\boldsymbol{X}\|_2 \ge \gamma\sqrt{n}$$

Then, for large n,

$$\begin{aligned}\frac{\mathsf{Vol}(B^n \cap \{|x_1| \le c/\sqrt{n}\})}{\mathsf{Vol}(B^n)} &\ge 1 - \mathbb{P}\left[\frac{|X_1|}{\|\boldsymbol{X}\|_2} > c/\sqrt{n}\right] \\ &\ge 1 - \mathbb{P}[|X_1| > \gamma c] - O(e^{-\Omega(n)}) = 1 - e^{-\Omega(c^2)}\end{aligned}$$

Again, most of the volume is both in the middle (equator) & thin shell?

Randomly selected unit vectors are likely orthogonal

Lemma Two randomly selected unit vectors are orthogonal with very hight probability.

Proof Pick two independent unit vectors from two slides ago

$$\boldsymbol{Y}^i = \left(\frac{X_1^i}{\|\boldsymbol{X}^i\|_2}, \dots, \frac{X_n^i}{\|\boldsymbol{X}^i\|_2} \right), \quad i = 1, 2.$$

Then, for any $\epsilon > 0$, as $n \to \infty$

$$\mathbb{P}[|\langle \boldsymbol{Y}^1, \boldsymbol{Y}^2 \rangle| > \epsilon] \approx \mathbb{P}[|\sum_j X_j^1 X_j^2| > \gamma \epsilon n] \to 0$$

What is the probability of this happening in low dimensions: 2 or 3?

Interpolation

Before we start looking into the NN generalization properties, let us consider the problem of interpolation: (see Section 5 in Pinkus (1999))

- Assume that σ is continuous ($\sigma \in C(\mathbb{R})$) and not a polynomial.
- Consider n data points $\{(x_i, y_i)\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$.
- Consider a shallow NN with one hidden layer and m neurons and weights $\{(w_j, b_j, a_j)\}_{j=1}^m, w_j \in \mathbb{R}^d, a_j, b_j \in \mathbb{R}$

$$f_m(x) = \sum_{j=1}^{m} a_j \sigma(w_j \cdot x - b_j)$$

- Interpolation problem: How many neurons m do we need to perfectly reproduce n data points, i.e., to have

$$f_m(x_1) = y_1, f_m(x_2) = y_2, \quad \cdots \quad , f_m(x_n) = y_n.$$

Interpolation

It turns out that it is enough to pick $m = n$ neurons.

Theorem Assume that σ is continuous ($\sigma \in C(\mathbb{R})$) and not a polynomial. For any set of distinct points $x_i \in \mathbb{R}^d$ and the associated $y_i, 1 \le i \le n$, there exist $\{(w_j, b_j, a_j)\}_{j=1}^m, w_j \in \mathbb{R}^d, a_j, b_j \in \mathbb{R}$, such that

$$\sum_{j=1}^{m} a_j \sigma(w_j \cdot x_i - b_j) = y_i, \quad 1 \le i \le n.$$

Proof:

- Assume first that σ is infinitely differentiable.
- Since $x_i \in \mathbb{R}^d$ are distinct, there exists $w \in \mathbb{R}^d$ such that $t_i = x_i \cdot w, 1 \le i \le n$ are all distinct.
  (This reduces the problem from d dimensions to 1.)
- Let $w_i = v_i w, v_i \in \mathbb{R}$. Now, the interpolation problem reduces to finding $a_i, v_i, b_i \in \mathbb{R}, 1 \le i \le n$ such that

$$\sum_{j=1}^{n} a_j \sigma(v_j t_i - b_j) = y_i, \quad 1 \le i \le n. \tag{1}$$

Interpolation

Proof:

- For Equation (1) to have a solution for any choice of $\{y_i\}$, it is enough that there exist $v_i, b_i \in \mathbb{R}, 1 \leq i \leq n$ such that

$$\det([\sigma(v_j t_i - b_j)]_{i,j=1}^n) \neq 0. \tag{2}$$

- On the other hand, if there are no $v_i, b_i, 1 \leq i \leq n$, then functions $\sigma(vt_i - b), 1 \leq i \leq n$ are linearly dependent for all $v, b \in \mathbb{R}$, i.e., there exist $\{c_i\}_{i=1}^n \not\equiv 0$ such that

$$\sum_{i=1}^n c_i \sigma(vt_i - b) = 0, \quad \text{for all } v, b \in \mathbb{R}. \tag{3}$$

- Next, since σ is infinitely differentiable and not a polynomial, there exist b_0 such that

$$\frac{d^j}{dv^j}\sigma(vt_i - b_0)\Big|_{v=0} = \sigma^{(j)}(-b_0)t_i^j, \quad \sigma^{(j)}(-b_0) \neq 0, 0 \leq j \leq n-1.$$

- Now, if we take the same derivatives as above in Equation (3) for $0 \leq j \leq n-1$, and then divide each of them by $\sigma^{(j)}(-b_0) \neq 0$

Interpolation

Proof:

- we obtain that constants $\{c_i\}_{i=1}^n$ must satisfy
$$\sum_{i=1}^n c_i t_i^j = 0, \quad 0 \le j \le n-1. \tag{4}$$
- The preceding system of equations has a unique solution since the following determinant, known as Vandermonde determinant, is non-zero (compute it explicitly for $n = 3$)
$$\det([t_i^{j-1}]_{i,j=1}^n) = \prod_{1 \le i < k \le n} (t_i - t_k) \ne 0$$
since t_i are distinct.
- Hence, the above implies that Equation (4) has a unique solution $\{c_i\}_{i=1}^n \equiv 0$, which contradicts our assumption in Equation (3), implying that the determinant in Equation (2) is not zero, i.e., Equation (1) has a solution. This proves the theorem for σ being infinitely differentiable.
- If σ is just continuous, we can always smooth it out by convolving it with infinitely differentiable function ϕ_δ (say Gaussian-like $\phi_\delta(t) = e^{-t^2/(2\delta^2)}/\sqrt{2\pi\delta}$, such that $\sigma_\delta(t) = \sigma * \phi_\delta(t)$ is infinitely differentiable and $|\sigma(t) - \sigma_\delta(t)| < \epsilon$; details omitted.

Interpolation and Overfitting

- Interpolation with ReLU, $\sigma(x) = x^+$, is even easier
  - Assume n data points $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}, 1 \leq i \leq n$
  - We can find a direction $\boldsymbol{v}$, such that the projections: $t_i = \boldsymbol{v} \cdot \boldsymbol{x}_i$ are all distinct.
  - Now, we have a one-dimensional problem $\{t_i, y_i\}_{i=1}^n$, which we can interpolate with $m = n$ neurons. We have seen this in the first lecture.
  - In fact, there are infinitely many interpolations with $m = n$ neurons.
- Can we interpolate n data points with $m < n$ neurons?
  - In general, the answer is no.
- Overfitting: The preceding interpolation problem show that any shallow NN with $m \geq n$ neurons can perfectly reproduce data, i.e., it can overfit.
  - The fact that, in many cases, training over-parametrized NN with $m \gg n$ does not lead to overfitting as often as one would expect is a mystery

Over-Parametrization

- Previous idea: **Composition of functions**
- New idea: **Over-parametrization**
  - Mathematically speaking: Passes the width $m \to \infty$ (this is the opposite of the previous work on expressiveness, where the goal was to minimize the $\#$ of neurons)
  - Mathematical tools: Laws of large numbers
    - "Smooth out" the layer functions
  - This idea was used in a number of recent papers to:
    - Connect NNs to Kernels
    - Shaw that all local minima are global
    - Shaw convergence to global minima
    - Shaw generalization bounds (recent, starting 2019+)
  - Based on the interpolation results, it is a mystery why these models generalize well/do not overfit.

Over-parametrized NNs and Kernels

Cho&Saul (2009) and Tsuchida et al. (2018), and others:

- Exploit large width, m, to explicitly compute the kernel corresponding to a hidden layer

Notation

- $\boldsymbol{w}_i$ - independent random initial weights with density $f(w), w \in \mathbb{R}^n$
- $\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ - hidden layer vector, m width of the hidden layer

$$\boldsymbol{h}(\boldsymbol{x}) := \frac{1}{\sqrt{m}}(\sigma(\langle \boldsymbol{w}_1, \boldsymbol{x}\rangle), \ldots \sigma(\langle \boldsymbol{w}_m, \boldsymbol{x}\rangle))$$

Then, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and large m

$$\langle \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{y})\rangle = \frac{1}{m}\sum_{i=1}^{m} \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}_i, \boldsymbol{y}\rangle) \tag{5}$$

$$\approx \int_{\mathbb{R}^n} \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{y}\rangle) f(\boldsymbol{w}) d\boldsymbol{w} =: k(\boldsymbol{x}, \boldsymbol{y}) \tag{6}$$

The "$\approx$" can be made precise in a probabilistic sense

Arc-Cosine Kernels

Theorem (Cho&Saul (2009)) Let $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1$, $\boldsymbol{w}$ be a vector of independent normal, $\mathcal{N}(0,1)$, random variables, and $\theta = \cos^{-1}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle)$. Then,

$$k(\boldsymbol{x},\boldsymbol{y}) = \frac{\pi - \theta}{2\pi}, \qquad \text{for } \sigma(x) = 1_{\{x \geq 0\}}$$
$$k(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\pi}(\sin\theta + (\pi - \theta)\cos\theta), \qquad \text{for } \sigma(x) = x_+$$

Proof Direct integration with respect to the Gaussian density: for step function $\sigma(x) = 1_{\{x \geq 0\}}$,

$$k(\boldsymbol{x},\boldsymbol{y}) = \int_{\boldsymbol{w} \in \mathbb{R}^d} 1_{\{\boldsymbol{x}\cdot\boldsymbol{w} \geq 0\}} 1_{\{\boldsymbol{y}\cdot\boldsymbol{w} \geq 0\}} f(\boldsymbol{w}) d\boldsymbol{w}, \quad f(\boldsymbol{w}) = \frac{e^{-\|\boldsymbol{w}\|_2^2/2}}{(2\pi)^{n/2}}$$

Key observation: due to rotational symmetry of the Gaussian density, we can pick the coordinates to align with $(\boldsymbol{x},\boldsymbol{y})$, i.e., $\boldsymbol{x} = (1,0,0,\ldots)$, and $\boldsymbol{y} = (y_1, y_2, 0, \ldots)$, and the preceding integral reduces to 2 dimensional:

$$k(\boldsymbol{x},\boldsymbol{y}) = \int_{(w_1,w_2)\in\mathbb{R}^2} 1_{\{w_1 \geq 0\}} 1_{\{y_1 w_1 + y_2 w_2 \geq 0\}} \frac{e^{-(w_1^2+w_2^2)/2}}{2\pi} dw_1 dw_2$$

Arc-Cosine Kernels

Proof Now, recall the norm $\|\boldsymbol{y}\|_2 = 1$, to obtain

$$k(\boldsymbol{x}, \boldsymbol{y}) = \int_{(w_1, w_2) \in \mathbb{R}^2} 1_{\{w_1 \geq 0\}} 1_{\{w_1 \cos\theta + w_2 \sin\theta \geq 0\}} \frac{e^{-(w_1^2 + w_2^2)/2}}{2\pi} dw_1 dw_2$$

where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Finally, write the preceding integral in polar coordinates: $w_1 = r\cos\phi, w_2 = r\cos\phi$, to obtain

$$\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{y}) &= \int_0^{2\pi} d\phi \int_0^{\infty} 1_{\{\cos\phi \geq 0\}} 1_{\{\cos\phi\cos\theta + \sin\phi\sin\theta \geq 0\}} \frac{e^{-r^2/2}}{2\pi} r dr \\
&= \frac{1}{2\pi} \int_0^{2\pi} 1_{\{\cos\phi \geq 0\}} 1_{\{\cos(\phi - \theta) \geq 0\}} d\phi \\
&= \frac{\pi - \theta}{2\pi};
\end{aligned}$$

draw a picture for the last equality.

Arc-Cosine Kernels: ReLU Case

Proof Again, recall the norm $\|\boldsymbol{y}\|_2 = 1$, to obtain

$$k(\boldsymbol{x}, \boldsymbol{y}) = \int w_1(w_1 \cos\theta + w_2 \sin\theta) 1_{\{w_1 \geq 0\}} 1_{\{w_1 \cos\theta + w_2 \sin\theta \geq 0\}} \frac{e^{-(w_1^2 + w_2^2)/2}}{2\pi} dw_1 dw_2$$

where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Finally, write the preceding integral in polar coordinates: $w_1 = r\cos\phi, w_2 = r\cos\phi$, to obtain

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{y}) &= \int_0^{2\pi} d\phi \int_0^{\infty} r^2 \cos\phi \cos(\phi - \theta) 1_{\{\cos\phi \geq 0\}} 1_{\{\cos\phi\cos\theta + \sin\phi\sin\theta \geq 0\}} \frac{e^{-r^2/2}}{2\pi} r dr \\ &= \frac{1}{\pi} \int_0^{2\pi} \frac{\cos\theta + \cos(2\phi - \theta)}{2} 1_{\{\cos\phi \geq 0\}} 1_{\{\cos(\phi - \theta) \geq 0\}} d\phi \\ &= \frac{1}{\pi} \int_\theta^{\pi} \frac{\cos\theta + \cos(2\phi - \theta)}{2} d\phi \\ &= \frac{\sin\theta + (\pi - \theta)\cos\theta}{2\pi}; \end{aligned}$$

draw a picture for the last equality.

Generalization to Rotationally Invariant Densities

Tsuchida et al. (2018)

- Rotationally invariant density $f(\boldsymbol{w})$: if for any $\boldsymbol{w}$ and orthogonal matrix R

  $$f(\boldsymbol{w}) = f(R\boldsymbol{w}) = f(\|\boldsymbol{w}\|_2)$$

  R is orthogonal if its raws and columns are orthogonal unit vectors
- Rotationally invariant distribution include: Gaussian distribution, the multivariate t-distribution, the symmetric multivariate Laplace distribution, and symmetric multivariate stable distributions.

Proposition 1 (Tsuchida et al. (2018)) The result of Cho&Saul (2009), which corresponds to ReLU, fully generalizes to rotationally invariant densities.

Proposition 4 (Tsuchida et al. (2018)) Extension to Leaky-ReLU for rotationally invariant densities.

Infinite Depth Networks: Degenerate Kernel

Corollary 8 (Tsuchida et al. (2018)) The normalized kernel converges to a degenerate fixed point at $\theta^* = 0$

Proof is based on contraction argument.

Additional comments

- This confirms the difficulty of trining very deep networks in view of the recent work on "Shattered Gradient Problem" by Balduzzi et al., 2017
- Check references in (Tsuchida et al. (2018)) for prior results of this type: Lee et al. (2017), Schoenholz et al. (2017), Poole et al. (2016) and Daniely (2016)

Over-Parametrization: No Bad Local Minima

Soudry and Carmon (2016): Probabilistic setup:

- Gaussian dropout noise $\mathcal{E}$ and leaky-ReLU like activations
- Data $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n], \boldsymbol{x}_i \in \mathbb{R}^{m_0}$ is smoothed by small Gaussian noise
- Mild over-parametrization: $m_0 m_1 \geq n$, where m_l is the width of the activation layer l (for interpolation results, we assume $m_1 \geq n$)

Theorem If $n \leq m_1 m_0$, then all differentiable local minima are global minima with MSE=0, $(\boldsymbol{X}, \mathcal{E})$ almost everywhere.

- They prove a similar result for the general depth
- Since they assume local minima to be differentiable, one can expect the convergence of GD locally
- However, there is no global convergence

Over-Parametrization: Convergence to Global Minimum

Today, we'll primarily focus on this recent paper:

- Gradient Descent Provably Optimizes Over-parameterized Neural Networks, by Du et al., ICLR, Feb 2019.
  - See the references therein and the follow up references for a comprehensive list on over-parametrization literature, e.g.:
    - No bad local minima: Data independent training error guarantees for multilayer neural networks, by Soudry and Carmon, 2016.
  - There are more recent papers that we'll cover later in the class. One of the attributes over-parametrization is good generalization error, e.g.
    - Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, by Arora et al., 2019.

Over-Parametrization: Convergence to Global Minimum

Some notational conventions:

- Get rid of the bias, b: augment $\boldsymbol{x}$ with an auxiliary feature $x_0 \equiv 1$ and call $w_0 = b$, then
$$b + \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \langle \boldsymbol{w}', \boldsymbol{x}' \rangle,$$
where $w' = (w_0, w_1, \ldots, w_d), x' = (x_0, x_1, \ldots, x_d)$. Hence, we get rid of b by embedding the problem in $d+1$ dimensions.
- For simplicity, data is often normalized on a hyper-sphere: $\|x\| = 1$
- For a one hidden layer NN, weights in the second layer can be simplified:

$$f(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \max(0, \langle \boldsymbol{w}_r, \boldsymbol{x} \rangle) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \frac{a_r}{|a_r|} \max(0, \langle |a_r| w_r, \boldsymbol{x} \rangle)$$

  - Hence, $|a_r|$ can be incorporate into random weights
  - a_r can be assumed Bernoulli $\{\pm 1\}$ since $a_r/|a_r| \in \{\pm 1\}$
  - Scaling $1/\sqrt{m}$ allows for Law of Large Numbers when computing kernels as $(1/\sqrt{m})^2 = 1/m$
  - $\boldsymbol{W}$ is $d \times m$ matrix

Quadratic Loss and Gradient Descent (GD)

- **Training**: Given data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n, \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. we minimize

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times m}} L(\boldsymbol{W})$$

for quadratic loss

$$L(\boldsymbol{W}) := \frac{1}{2} \sum_{i=1}^n (f(\boldsymbol{x}_i) - y_i)^2$$

- **Gradient:**

$$\frac{\partial L(\boldsymbol{W})}{\partial \boldsymbol{w}_r} = \frac{1}{\sqrt{m}} \sum_{i=1}^n (f(\boldsymbol{x}_i) - y_i) a_r \boldsymbol{x}_i 1_{\{\langle \boldsymbol{w}_r, \boldsymbol{x}_i \rangle \geq 0\}}$$

- **Gradient Descent (GD):** for step size $\eta > 0$

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) - \eta \frac{\partial L(\boldsymbol{W}(k))}{\partial \boldsymbol{W}(k)}$$

Continuous Time Approximation

- **Gradient Flow:** Recall from the beginning of the class, gradient descent with infinitesimal step size

$$\frac{\boldsymbol{W}(k+1) - \boldsymbol{W}(k)}{\eta} \approx \frac{d\boldsymbol{W}(t)}{dt} = -\frac{\partial L(\boldsymbol{W}(t))}{\partial \boldsymbol{W}(t)}, \tag{7}$$

  where $\boldsymbol{W}(t) \in \mathbb{R}^{d\times m}$ is the continuous flow of gradient descednt.

- Denote the prediction on input $\boldsymbol{x}_i$ at time t as

$$u_i(t) = f(W(t), \boldsymbol{a}, \boldsymbol{x}_i), \quad i = 1, \ldots, n$$

  and

$$\boldsymbol{u}(t) = (u_1(t), \ldots, u_n(t))$$

Dynamical System

Time dynamics of each prediction

$$\begin{aligned}\frac{du_i(t)}{dt} &= \sum_{r=1}^{m} \left\langle \frac{\partial f(\boldsymbol{W}(t), \boldsymbol{x}_i)}{\partial \boldsymbol{w}_r(t)}, \frac{d\boldsymbol{w}_r(t)}{dt} \right\rangle \\ &= \sum_{j=1}^{n} (y_j - u_j) \sum_{r=1}^{m} \left\langle \frac{\partial f(\boldsymbol{W}(t), \boldsymbol{x}_i)}{\partial \boldsymbol{w}_r(t)}, \frac{\partial f(\boldsymbol{W}(t), \boldsymbol{x}_j)}{\partial \boldsymbol{w}_r(t)} \right\rangle \\ &=: \sum_{j=1}^{n} (y_j - u_j) H_{ij}(t)\end{aligned}$$

where in the second equality we used the gradient flow equation (7) and

$$H_{ij}(t) = \frac{1}{m} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \sum_{r=1}^{m} 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}_r(t)\rangle \geq 0, \langle \boldsymbol{x}_i, \boldsymbol{w}_r(t)\rangle \geq 0\}}$$

Define $n \times n$ matrix $\boldsymbol{H}(t) = \{H_{ij}(t)\}$

Dynamical System

The vector of predictors, $\boldsymbol{u}(t)$, evolves as

$$\frac{d}{dt}\boldsymbol{u}(t) = \boldsymbol{H}(t)(\boldsymbol{y} - \boldsymbol{u}(t)) \tag{8}$$

Objective: Prove that

$$\boldsymbol{u}(t) \to \boldsymbol{y}, \quad \text{as } t \to \infty$$

for any

- initial random weights $\boldsymbol{w}_r = \boldsymbol{w}_r(0)$
- finite set of data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$

Key Ideas

- **Over-parametrization - large width** m, by the Laws of Large Numbers

$$H_{ij}(0) = \frac{1}{m}\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \sum_{r=1}^{m} 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}_r(0)\rangle \geq 0, \langle \boldsymbol{x}_j, \boldsymbol{w}_r(0) \geq 0\rangle\}}$$

$$\approx \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \, \mathbb{E}\, 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}_r(0)\rangle \geq 0, \langle \boldsymbol{x}_j, \boldsymbol{w}_r(0) \geq 0\rangle\}} =: H_{ij}^{\infty}$$

- Assume m so large, $m = \Omega(n^6)$, so that the initial random state does not change much during training, $\boldsymbol{W}(t) \approx \boldsymbol{W}$, and therefore

$$\boldsymbol{H}(t) \approx \boldsymbol{H}(0) \approx \boldsymbol{H}^{\infty}$$

- Hence, the vector of predictors, $\boldsymbol{u}(t)$, evolves as

$$\frac{d}{dt}\boldsymbol{u}(t) = \boldsymbol{H}(t)(\boldsymbol{y} - \boldsymbol{u}(t)) \approx \boldsymbol{H}^{\infty}(\boldsymbol{y} - \boldsymbol{u}(t))$$

linear system with constant (time invariant) coefficients

Key Ideas: Minimum Eigenvalue

For the convergence of the linear system

$$\frac{d}{dt}\boldsymbol{u}(t) \approx \boldsymbol{H}^{\infty}(\boldsymbol{y} - \boldsymbol{u}(t))$$

- One needs the minimum eigenvalue

  $$\lambda_0 := \lambda_{\min}(\boldsymbol{H}^{\infty}) > 0$$

  which is assumed in the paper. By Cho&Saul (2009) result that we just proved

  $$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathbb{E}\, 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}_r) \rangle \geq 0, \langle \boldsymbol{x}_j, \boldsymbol{w}_r \geq 0 \rangle\}} = \frac{1}{2} - \frac{1}{2\pi} \cos^{-1}\left(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle\right))$$

  Hence, one should have $\lambda_0 := \lambda_{\min}(\boldsymbol{H}^{\infty}) > 0$ if there is no parallel pair $\boldsymbol{x}_i, \boldsymbol{x}_j$

Formal Theorem

Assumptions:

- Positive minimum eigenvalue: $\lambda_0 := \lambda_{\min}(\boldsymbol{H}^\infty) > 0$
- Scaled data: $\|\boldsymbol{x}_i\|_2 = 1$ and $|y_i| \leq C$
- Over-parametrization:
$$m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$$
- Random initialization: $\boldsymbol{w}_r \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $a_r \sim \text{Uniform}(\{\pm 1\})$

Theorem 3.2 Then, on a set of probability at least $1 - \delta$

$$\|\boldsymbol{u}(t) - \boldsymbol{y}\|_2^2 \leq \exp(-\lambda_0 t)\|\boldsymbol{u}(0) - \boldsymbol{y}\|_2^2$$

Proof Sketches

$\boldsymbol{H}(0)$ and $\boldsymbol{H}^\infty$ are close as expected by the Law of Large Numbers

Lemma 1 If $m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$, then with probability at least $1-\delta$,

$$\|\boldsymbol{H}(0) - \boldsymbol{H}^\infty\| \leq \frac{\lambda_0}{4} \quad \text{and} \quad \lambda_{\min} \geq \frac{3}{4}\lambda_0$$

Proof: For each (i,j), by Hoeffding inequality (proved later), with probability $1-\delta'$,

$$|\boldsymbol{H}_{ij}(0) - \boldsymbol{H}_{ij}^\infty| \leq \frac{2\sqrt{\log(1/\delta')}}{\sqrt{m}}$$

Setting $\delta' = n^2\delta$ and using the union bound one obtains for all (i,j) with probability at least $1-\delta$ that

$$|\boldsymbol{H}_{ij}(0) - \boldsymbol{H}_{ij}^\infty| \leq \frac{4\sqrt{\log(n/\delta)}}{\sqrt{m}}$$

implying (recall $m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$)

$$\|\boldsymbol{H}(0) - \boldsymbol{H}^\infty\|_2^2 \leq \|\boldsymbol{H}(0) - \boldsymbol{H}^\infty\|_F^2 \leq \sum_{i,j} |\boldsymbol{H}_{ij}(0) - \boldsymbol{H}_{ij}^\infty|^2 \leq \frac{16 n^2 \log^2(n/\delta)}{m}$$

Proof Sketches

If weights $\boldsymbol{w}_r(t)$ don't move much, then $\boldsymbol{H}(t)$ is close to $\boldsymbol{H}^\infty$

Lemma 2 If $\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\| \leq \frac{c\lambda_0}{n^2} =: R$, then with probability at least $1-\delta$, $\lambda_{\min}(\boldsymbol{H}(t)) > \lambda_0/2$

Proof: Define event

$$A_{ir} = \{\boldsymbol{w} : \|\boldsymbol{w} - \boldsymbol{w}_r(0)\| \leq R, 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}_r(0)\rangle \geq 0\}} \neq 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}\rangle \geq 0\}}\}$$

Not that this event happens iff $|\langle \boldsymbol{x}_i, \boldsymbol{w}_r(0)\rangle| < R$
Also, recall $\boldsymbol{w}_r(0) \sim N(\boldsymbol{0}, \boldsymbol{I})$. Hence, by anti-concentration inequality for Gaussian

$$\mathbb{P}[A_{ir}] = \mathbb{P}_{Z\sim N(0,1)}[|Z| < R] \leq \frac{2R}{\sqrt{2\pi}}$$

Hence, we can bound entry-wise deviation of $\boldsymbol{H}(t)$

$$\mathbb{E}[|H_{ij}(t) - H_{ij}(0)|] = \mathbb{E}\left[\frac{1}{m}\left|\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \sum_{r=1}^m 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}_r(t)\rangle \geq 0, \langle \boldsymbol{x}_j, \boldsymbol{w}_r(t) \geq 0\rangle\}} - \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \sum_{r=1}^m 1_{\{\langle \boldsymbol{x}_i, \boldsymbol{w}_r(0)\rangle \geq 0, \langle \boldsymbol{x}_j, \boldsymbol{w}_r(0) \geq 0\rangle\}}\right|\right]$$

Proof Sketches

Proof: Implying

$$\mathbb{E}[|H_{ij}(t) - H_{ij}(0)|] \leq \frac{1}{m} \sum_r \mathbb{P}[A_{ir} \cup A_{jr}] \leq \frac{4R}{\sqrt{2\pi}}$$

$$\|\boldsymbol{H}(0) - \boldsymbol{H}^\infty\|_2 \leq \|\boldsymbol{H}(0) - \boldsymbol{H}^\infty\|_F \leq \sqrt{\sum_{i,j} |\boldsymbol{H}_{ij}(0) - \boldsymbol{H}_{ij}^\infty|^2} \leq Cn^2 R$$

Therefore, using $R = \frac{c\lambda_0}{n^2}$, the lower eigenvalue can be bounded

$$\lambda_{\min}(\boldsymbol{H}(t)) \geq \lambda_{\min}(\boldsymbol{H}(0)) - Cn^2 R \geq \frac{\lambda_0}{2}$$

Proof Sketches

$\boldsymbol{w}_r(t)$ don't move much from the initial value $\boldsymbol{w}_r(0)$

Lemma 3 Assume for $0 \le s \le t$, $\lambda_{\min}(\boldsymbol{H}(s)) \ge \lambda_0/2$. Then, $\|\boldsymbol{y} - \boldsymbol{u}(t)\|_2^2 \le \exp(-\lambda_0 t)\|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^2$ and

$$\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2 \le \frac{\sqrt{n}\|\boldsymbol{y} - \boldsymbol{u}(0)\|_2}{\sqrt{m}\lambda_0} =: R'$$

Proof: We can write the dynamics for the norm

$$\frac{d}{dt}\|\boldsymbol{y} - \boldsymbol{u}(t)\|_2^2 = -2(\boldsymbol{y} - \boldsymbol{u}(t))^\top \boldsymbol{H}(t)(\boldsymbol{y} - \boldsymbol{u}(t)) \le -\lambda_0\|\boldsymbol{y} - \boldsymbol{u}(t)\|_2^2$$

which implies $\|\boldsymbol{y} - \boldsymbol{u}(t)\|_2^2 \le \exp(-\lambda_0 t)\|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^2$, meaning that $\boldsymbol{u}(t) \to \boldsymbol{y}$ exponentially fast.

Proof Sketches

Proof: Next, for $0 \leq s \leq t$

$$
\begin{aligned}
\left\| \frac{d}{ds} \boldsymbol{w}_r(s) \right\| &= \left\| \sum_{i=1}^{n} (y_i - u_i(s)) \frac{1}{\sqrt{m}} a_r \boldsymbol{x}_i 1_{\{\langle \boldsymbol{w}_r(s), \boldsymbol{x}_i \rangle \geq 0\}} \right\| \\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{n} |y_i - u_i(s)| \leq \frac{\sqrt{n}}{\sqrt{m}} \|\boldsymbol{y} - \boldsymbol{u}(s)\|_2 \\
&\leq \frac{\sqrt{n}}{\sqrt{m}} \exp(-\lambda_0 s) \|\boldsymbol{y} - \boldsymbol{u}(0)\|_2
\end{aligned}
$$

Hence, integrating the preceding derivative

$$
\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2 \leq \int_0^t \left\| \frac{d}{ds} \boldsymbol{w}_r(s) \right\| ds \leq \frac{\sqrt{n} \|\boldsymbol{y} - \boldsymbol{u}(0)\|_2}{\sqrt{m} \lambda_0}
$$

Proof Sketches

Finally, we need to show that, if $R' < R$, the conditions in Lemma 2 and 3 hold for all $t \geq 0$.

Lemma 4 If $R' < R$, we have for all $t \geq 0, \lambda_{\min}(H(t)) \geq \lambda_0/2$, for all r, $\|\boldsymbol{w}_r(t) - w_r(0)\|_2 \leq R'$ and $\|y - u(t)\|_2^2 \leq \exp(-\lambda_0 t)\|y - u(0)\|_2^2$.

Proof: The proof is by contradiction: Suppose the conclusion does not hold at time t. Hence, there exists r, such that $\|\boldsymbol{w}_r(t) - w_r(0)\|_2 > R'$ or $\|y - u(t)\|_2^2 > \exp(-\lambda_0 t)\|y - u(0)\|_2^2$.

Then, by Lemma 3, there exists $s \leq t$, such that $\lambda_{\min}(\boldsymbol{H}(s)) < \lambda_0/2$.

Next, by Lemma 2, there exists

$$t_0 = \inf\{t \geq 0 : \max_r \|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2^2 \geq R\}$$

Proof Sketches

Proof: Thus, at t_0, there exists r, such that

$$\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2^2 = R.$$

By Lemma 2, we know that $\lambda_{\min} H(t) \geq \lambda_0/2$ for $t \leq t_0$.
However, by Lemma 3, we know that $\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2^2 < R' < R$, which is a contradiction.
For the other case, at time t, $\lambda_{\min}(\boldsymbol{H}(t)) < \lambda_0$, we know there exists

$$t_0 = \inf\{t \geq 0 : \max_r \|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2^2 \geq R\}$$

The rest of the proof is the same as in the previous case.

Proof of the Theorem

Hence, in view of Lemmas 1-4, it is enough to show that $R' < R$, which is equivalent to

$$m = \Omega\left(\frac{n^5\|y - u(0)\|_2^2}{\lambda_0^4\delta^2}\right).$$

For $\|y - u(0)\|_2^2$, note that

$$\begin{aligned}\mathbb{E}[\|y - u(0)\|_2^2] &= \sum_{i=1}^{n}(y_i^2 - 2y_i\,\mathbb{E}[f(\boldsymbol{W}(0), \boldsymbol{a}, \boldsymbol{x}_i)] + \mathbb{E}[f(\boldsymbol{W}(0), \boldsymbol{a}, \boldsymbol{x}_i)^2] \\ &= \sum_{i=1}^{n}(y_i^2 + 1) = O(n).\end{aligned}$$

Thus, by Markov's inequality, with probability at least $1 - \delta$, we have $\|y - u(0)\|_2^2 = O(n/\delta)$

Therefore, the following choice of m satisfies all the consitions

$$m = \Omega\left(\frac{n^6}{\lambda_0^4\delta^3}\right).$$

Over-parametrization and Generalization

Followup work:
Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, by Arora et al., 2019.

- Refine the analysis of Du et al. (2019), but assume even wither networks - see Theorem 4.1

$$m = \Omega\left(\frac{n^7}{\lambda_0^4 \kappa^2 \delta^4 \epsilon^2}\right).$$

- Prove a generalization bound, in Theorem 5.1, that doesn't depend on m.
  Is this surprising?

Useful Tools in ML: Concentration Inequalities

Hoeffding's Bound: used in Du et al. proof.

Lemma Let Z_i be i.i.d. random variables with $\mathbb{E}\, Z_i = \mu$ and bounded support $\mathbb{P}[a \leq Z_i \leq b] = 1$. Then, for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$$

Proof: Let's center Z_i: $X_i = Z_i - \mu$ and set $\bar{X} = (1/m)\sum X_i$. Then, for $\epsilon > 0$ and $\lambda > 0$, by Markov's inequality and i.i.d.

$$\mathbb{P}[\bar{X} \geq \epsilon] = \mathbb{P}[e^{\lambda \bar{X}} \geq e^{\lambda\epsilon}] \leq e^{-\lambda\epsilon}\,\mathbb{E}[e^{\lambda \bar{X}}] = e^{-\lambda\epsilon}\left(\mathbb{E}[e^{\lambda X_1/m}]\right)^m \tag{9}$$

Next, we show that

$$\mathbb{E}[e^{\lambda X_1}] \leq e^{\lambda^2 (b-a)^2/8} \tag{10}$$

Let $a' = a - \mu$, $b' = b - \mu$ and note that $b' - a' = b - a$. By convexity

$$\mathbb{E}[e^{\lambda X_1}] \leq \frac{b' - \mathbb{E}[X_1]}{b-a} e^{\lambda a'} + \frac{\mathbb{E}[X] - a'}{b-a} e^{\lambda b'} = \frac{b'}{b-a} e^{\lambda a'} - \frac{a'}{b-a} e^{\lambda b'} =: f(\lambda)$$

Useful Tools in ML: Concentration Inequalities

Proof: Next, if $h = \lambda(b-a)$ and $p = -a'/(b-a)$, then, by Taylor's theorem

$$L(h) := \log(f(h/(b-a))) = -hp + \log(1 - p + pe^h) \leq \frac{h^2}{8}$$

since $L(0) = L'(0) = 0$ and $L''(h) \leq 1/4$ for all h; this proves (10). Now, using (10) in (9), we obtain

$$\mathbb{P}[\bar{X} \geq \epsilon] \leq e^{-\lambda\epsilon}\,\mathbb{E}[e^{\lambda\bar{X}}] = e^{-\lambda\epsilon}\left(\mathbb{E}[e^{\lambda X_1/m}]\right)^m \leq e^{-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m}},$$

which is minimized for

$$\lambda^* = \frac{4me}{(b-a)^2}$$

This concludes the proof of Hoeffding's bound.

We'll cover more concentration inequalities in the future.

Useful Tools: Eigendecomposition

Since Gram matrix, H, is positive definite, it can be decomposed as

$$H = U^{-1}\Lambda U,$$

where Λ is diagonal matrix with $\Lambda_{ii} = \lambda_i$ (and U is unitary matrix, i.e., $U^{-1} = U^*$ - conjugate transpose of U)

This was used in the proof of Lemma 3 of Du et al.

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{y} - \boldsymbol{u}(t)\|_2^2 &= -2(\boldsymbol{y} - \boldsymbol{u}(t))^\top \boldsymbol{H}(t)(\boldsymbol{y} - \boldsymbol{u}(t)) \\
&= -2(\boldsymbol{y} - \boldsymbol{u}(t))^\top \boldsymbol{U}^{-1}\boldsymbol{\Lambda}\boldsymbol{U}(\boldsymbol{y} - \boldsymbol{u}(t)) \\
&\leq -2\lambda_{\min}\|\boldsymbol{y} - \boldsymbol{u}(t)\|_2^2
\end{aligned}$$

Useful Tools: Matrix Norms

L_2 norm

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$$

For symmetric/hermitian matrices

$$\|A\|_2 = \rho(A) \geq |\lambda_{\min}|$$

where $\rho(A) = \max\{|\lambda_i\|\}$ is the spectral radius.
Frobenius norm

$$\|A\|_F^2 = \sum_{ij} (a_{ij})^2 = \mathsf{trace}(A^\top A)$$

It bounds the L_2 norm

$$\|A\|_2 \leq \|A\|_F$$

Matrix Perturbation Theory: Hoffman-Wielandt Inequality

A matrix A is normal iff $AA^* = A^*A$, where A^* is conjugate transpose of A. See Theorem 6.3.5 in Chapter 6, Matrix Analysis, by Horn & Johnson, 2013.

Theorem Let A and its perturbation $A+E$ be normal matrices with respective eigenvalues λ_i and $\hat{\lambda}_i$. Then there exists a permutation σ such that

$$\sum_i |\hat{\lambda}_{\sigma(i)} - \lambda_i|^2 \leq \|E\|_F^2$$

Proof: Normal matrices are diagonalizable by unitary matrices. Let $A = U\Lambda U^*$ and $A + E = V\hat{\Lambda}V^*$. Then, using the unitary invariance of Frobenius norm,

$$\begin{aligned}
\|E\|_F^2 &= \|V\hat{\Lambda}V^* - U\Lambda U^*\|_F^2 \\
&= \|U^*V\hat{\Lambda} - \Lambda U^*V\|_F^2 \\
&= \|W\hat{\Lambda} - \Lambda W\|_F^2 = \sum_i |\hat{\lambda}_i - \lambda_i|^2 |w_{ij}|^2
\end{aligned}$$

where $|w_{ij}|^2$ doubly stochastic.

Matrix Perturbation Theory: Hoffman-Wielandt Inequality

Proof: (continued) Then

$$\|E\|_F^2 = \|W\hat{\Lambda} - \Lambda W\|_F^2 = \sum_i |\hat{\lambda}_i - \lambda_i|^2 |w_{ij}|^2$$
$$\geq \min_{s_{ij}} \sum_i |\hat{\lambda}_i - \lambda_i|^2 s_{ij}$$

where $\{s_{ij}\}$ is any doubly stochastic matrix. Hence, by Birkhoff-von Neumann theorem, the preceding minim is achieved when $\{s_{ij}\}$ is a permutation matrix, which concludes the proof.

Corollary If A and $A + E$ are real symmetric (Hermitian) with their eigenvalues ordered $\lambda_1 \leq \cdots \leq \lambda_n$ and $\hat{\lambda}_1 \leq \cdots \leq \hat{\lambda}_n$, then σ can be chosen to be an identity, i.e.,

$$\sum_i (\hat{\lambda}_i - \lambda_i)^2 \leq \|E\|_F^2$$

Reading

- Local vs global minima and GD convergence
  - Gradient Descent Provably Optimizes Over-parameterized Neural Networks, by Du et al., ICLR, Feb 2019.
  - Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, by Arora et al., 2019.
  - No bad local minima: Data independent training error guarantees for multilayer neural networks, by Soudry and Carmon, 2016.
  - See the references in the preceding papers and the follow up ones on Google Scholar
- Connection between NNs and Kernels
  - Invariance of Weight Distributions in Rectified MLPs, by Tsuchida et al., Jun 2018.
  - Kernel Methods for Deep Learning, by Cho & Saul, 2009.
  - Check the reference lists in these papers for additional readings
- New Developing Monograph
  - Deep learning theory lecture notes, by Telgarsky, Oct 2021.

Have Fun!