Geometry of Deep Polynomial Neural Network

Maksym Zubkov

Math and ML Reading Group

February 16, 2024

Set Up

- Let X be a collection of points in $\mathbb{R}^{n_1}$ and Y be a collection of points in $\mathbb{R}^{n_h}$ i.e.

$$X = \{x_1, \ldots, x_k\} \text{ and } Y = \{y_1, \ldots, y_k\}.$$

- Ideally, we want to find some continuous function $f \in C(\mathbb{R}^{n_1}, \mathbb{R}^{n_h})$ s.t.

$$f(x_i) = y_i \text{ for all } i.$$

- By **a model space** $\mathcal{M}$, we will call a space of continuous functions $C(\mathbb{R}^{n_1}, \mathbb{R}^{n_h})$.
- *How can we find such f?*

In other settings, we can have other model spaces. For example, *probability distributions*.

What's a Neural Network?

- For that, consider a new map $p_\theta : \mathbb{R}^{n_1} \to \mathbb{R}^{n_h}$ that consists of a composition of affine linear transformation W_i with a non-linear function σ

$$p_\theta : \mathbb{R}^{n_1} \xrightarrow{W_1} \mathbb{R}^{n_2} \xrightarrow{W_2} \mathbb{R}^{n_3} \to \cdots \to \mathbb{R}^{n_{k-1}} \xrightarrow{W_h} \mathbb{R}^{n_h}$$

$$p_\theta(\mathbf{x}) = W_h \sigma W_{h-1} \sigma \ldots W_2 \sigma W_1 \mathbf{x}$$

  where $W_i \mathbf{x} = A_i \mathbf{x} + b_i$ with A_i being a linear transformation $\mathbb{R}^{n_i} \to \mathbb{R}^{n_{i+1}}$ and b_i being a vector in $\mathbb{R}^{n_{i+1}}$
- We can see that p_θ lives in a space of continuous functions from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_h}$ i.e. $p_\theta \in C(\mathbb{R}^{n_1}, \mathbb{R}^{n_h})$.
- p_θ is **a neural network** *(NN)*.

NN Architecture

- Now, f and p_θ live in the same *Model space* $C(\mathbb{R}^{n_1}, \mathbb{R}^{n_h})$.
- Let's collect all A_i and b_i into a set

$$\theta = \{(A_i, b_i) \in \mathbb{R}^N\}$$

  where N is a number of parameters in A_i and b_i.
- The space $\theta = \mathbb{R}^N$ is called **a parameter space** $\mathcal{P}$.
- Let $\mathbf{n} = (n_1, n_2, \ldots, n_h)$. We will call a tuple $(\mathbf{n}, \sigma)$ to be **an architecture** of a NN p_θ.
- In the literature, A_i are called **weights** and b_i are called **biases**.

Objects

Training data set:	(X, Y)
NN:	p_θ
Affine Linear Transformation:	$W_i\mathbf{x} = A_i\mathbf{x} + b_i$
Activation function:	σ
Weights:	$\theta = (A_i, b_i)$
Model Space:	$C(\mathbb{R}^{n_1}, \mathbb{R}^{n_h})$
Parameter Space	$\mathbb{R}^N$, N is the number of weights. **NN:**

Weight Map

- Next, let's define a weight map

$$\Psi : \mathcal{P} \to \mathcal{M}$$

$$\theta \mapsto p_\theta$$

- If we have a notion of a distance (metric) $\|\cdot\|$, then we can define **a loss function**

$$loss(p_\theta, (X, Y)) = \sum_{i=1}^{k} \|p_\theta(x_i) - y_i\|$$

- Usually, when we initialize initial random weights θ, the *loss* is pretty big. The goal is adjust our weights via gradient descent in $\mathcal{M}$ to minimize the loss function

Further Questions and Concepts to Learn

- Universal Approximation Theorem (why can we even do it?)
- Over fitting (ability to generalize NN)
- Getting stuck in local minima (a loss function landscape)
- Best initialization
- Way to optimize (Stochastic Gradient Descent, Adam optimizer)
- Different models $\mathcal{M}$ require different NN architectures.

What are Deep Polynomial Neural Networks (DPNNs)?

A PNN is defined as follows:

- It's NN without bias i.e. $\theta = (A_i, 0)$.
- It's activation function $\sigma := \rho_r$ is given by a monomial x^r i.e. ρ_r is defined by the entrywise operation

$$\rho_r(\mathbf{x}) = (x_1^r, \ldots, x_n^r).$$

- Thus the DPNN outputs for each coordinate a homogeneous polynomials i.e.

$$p_\theta(\mathbf{x}) = (p_\theta^1(\mathbf{x}), \ldots, p_\theta^{n_h}(\mathbf{x}))$$

- The model space $\mathcal{M}$ is given by a product of symmetric spaces $(\mathrm{Sym}_{r^{h-1}}(\mathbb{R}^{n_1}))^{n_h}$ i.e. $\mathrm{Sym}_{r^{h-1}}(\mathbb{R}^{n_1})$ is a space of homogeneous polynomial of degree r^{h-1} in n_1 variables.

Polynomial Neural Network — Example

This PNN has *architecture* $d = (3, 2, 1), r = 2$, and is given by the polynomial map

$$p_\theta : \mathbb{R}^3 \to \mathbb{R}^1, \mathbf{x} \mapsto W_2 \rho_2 W_1 \mathbf{x}$$

Here we have:

- ρ_2 is the activation function that squares each coordinate.
- W_1 and W_2 are linear transformations.

Parameter Map

We can compute the polynomial $p_\theta(\mathbf{x})$:

$$p_\theta(\mathbf{x}) = (W_2\rho_2W_1)\mathbf{x} = \begin{pmatrix} b_1 & b_2 \end{pmatrix} \rho_2 \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} =$$

$$= \begin{pmatrix} b_1 & b_2 \end{pmatrix} \begin{pmatrix} (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)^2 \\ (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)^2 \end{pmatrix} = b_1q_1^2 + b_2q_2^2$$

where $q_i := a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3$.

$$\Psi : \mathbb{R}^8 \to \mathrm{Sym}_2(\mathbb{R}^3) \cong \mathbb{R}^6$$

$$(a_{ij}, b_k)_{i,j,k} \mapsto p_\theta(x) = b_1(a_{11}x_1 + a_{12}x_2 + a_{13}x_3)^2 + b_2(a_{21}x_1 + a_{22}x_2 + a_{23}x_3)^2$$

Example $d = (3, 2, 1), \; r = 2$

For architecture $d = (3, 2, 1)$, $r = 2$ and parameters

$$\theta = \left[W_1 = \begin{pmatrix} b_1 & b_2 \end{pmatrix}, \; W_2 = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \right],$$

the resulting map Ψ is given by

$$\theta \mapsto \begin{pmatrix} b_1 a_{11}^2 + b_2 a_{21}^2 \\ b_1 a_{12}^2 + b_2 a_{22}^2 \\ b_1 a_{13}^2 + b_2 a_{23}^2 \\ 2(b_1 a_{11} a_{12} + b_2 a_{21} a_{22}) \\ 2(b_1 a_{11} a_{13} + b_2 a_{21} a_{23}) \\ 2(b_1 a_{12} a_{13} + b_2 a_{22} a_{23}) \end{pmatrix}$$

with the entries that are resulting coefficients of a homogeneous polynomial $b_1 q_1^2 + b_2 q_2^2$.

Neuromanifolds and -varieties

$\mathcal{M} := \mathrm{Im}(\Psi)$ denotes the *neuromanifold*. This is a semialgebraic set.

Its Zariski closure $\mathcal{V} = \overline{\mathcal{M}}$ is called the *neurovariety*.

An architecture of NN is *filling* if $\mathcal{V} = (\mathrm{Sym}_{r^{h-1}}(\mathbb{R}^{n_1}))^{n_h}$. In this case, we say that $\mathcal{M}$ is *thick*.

Question: What architectures are filling?

Single Output Networks

Next, let's consider networks with the architecture $d = (n, m, 1)$ for any $r \in \mathbb{N}$.
Then $p_\theta \in \mathrm{Sym}_r(\mathbb{R}^n)$ as

$$p_\theta(x) = b_1 q_1(x)^r + b_2 q_2(x)^r + \cdots + b_m q_m(x)^r \text{ with}$$

$$q_i(x) = a_{i1}x_1 + \cdots + a_{in}x_n, \; i = 1, 2 \ldots, m$$

So, we can see that

$$\mathcal{M}_{d,r} = \{p_\theta \in \mathrm{Sym}_r(\mathbb{R}^n) \mid p_\theta = b_1 q_1^r + b_2 q_2^r + \cdots + b_m q_m^r\}$$

Single Output Networks: $d = (n, m, 1)$ and $r = 2$

The neuromanifold $\mathcal{M}_{d,2} \subseteq \mathrm{Sym}_2(\mathbb{R}^n)$ is given by $b_1 q_1^2 + b_2 q_2^2 + \cdots + b_m q_m^2$.

Question: When is $\mathcal{M}_{d,2} = \mathrm{Sym}_2(\mathbb{R}^n)$?

- Take some $Q \in \mathrm{Sym}_2(\mathbb{R}^n)$.
- To each Q there's a corresponding symmetric matrix A of size $n \times n$.
- Then we can see that $Q = b_1 q_1^2 + b_2 q_2^2 + \cdots + b_m q_m^2$ if and only if

$$A = b_1 v_1^T v_1 + b_2 v_2^T v_2 + \cdots + b_m v_m^T v_m$$

  for some row vectors $v_i, \; i = 1, \ldots, m$.

So, $\mathcal{M}_{d,2}$ is described by symmetric matrices of rank at most m.

Single Output Networks: $d = (n, m, 1)$ and $r = 2$

- $\mathcal{M}_{d,2} = \mathrm{Sym}_2(\mathbb{R}^n)$ for $m \geq n$ as we need exactly n linear terms to hit the full rank of any symmetric matrix.
- $\mathcal{M}_{d,2} = \mathcal{V}_{d,2} \subsetneq \mathrm{Sym}_2(\mathbb{R}^n)$ for $m < n$. The image is given by symmetric matrices of rank $\leq m$. In other words, the image is cut out by $(m+1) \times (m+1)$ minors.

Example: Recall $d = (3, 2, 1)$, $r = 2$.

Then $p_\theta \in \mathcal{M}_{d,2}$ if and only if $\det(A) = 0$ where $A = b_1 v_1^T v_1 + b_2 v_2^T v_2 =$

$$= \begin{pmatrix} b_1 a_{11}^2 + b_2 a_{21}^2 & 2(b_1 a_{11} a_{12} + b_2 a_{21} a_{22}) & 2(b_1 a_{11} a_{13} + b_2 a_{21} a_{23}) \\ 2(b_1 a_{11} a_{12} + b_2 a_{21} a_{22}) & b_1 a_{12}^2 + b_2 a_{22}^2 & 2(b_1 a_{12} a_{13} + b_2 a_{22} a_{23}) \\ 2(b_1 a_{11} a_{13} + b_2 a_{21} a_{23}) & 2(b_1 a_{12} a_{13} + b_2 a_{22} a_{23}) & b_1 a_{13}^2 + b_2 a_{23}^2 \end{pmatrix}.$$

Single Output Networks: $d = (m, n, 1)$ and $r > 2$

The neuromanifold $\mathcal{M}_{d,r} \subset \mathrm{Sym}_r(\mathbb{R}^n)$ is given by $b_1 q_1^r + b_2 q_2^r + \cdots + b_m q_m^r$.

- Instead of a symmetric matrix A, we have a symmetric tensor T.
- Instead of $A = b_1 v_1^T v_1 + b_2 v_2^T v_2 + \cdots + b_m v_m^T v_m$, we have

$$T = b_1 v_1^{\otimes r} + b_2 v_2^{\otimes r} + \cdots + b_m v_m^{\otimes r}$$

- Unfortunately, the set of tensors with rank $\leq r$ is not closed.
- So, understanding $\mathcal{M}_{d,r}$ is equivalent to understanding the set of real symmetric tensors T of "some" symmetric rank m

Example: $d = (3, m, 1)$ and $r = 3$

Take a homogeneous polynomial f of degree 3 in 3 variables x, y, and z. According to [?], we can find a change of basis with real coefficients s.t.

$$f(x, y, z) \mapsto g(x, y, z) = x^3 + y^3 + z^3 + \lambda xyz \text{ with } \lambda \in \mathbb{R}.$$

We know the following about the symmetric tensor T_g

- ▷ if $\lambda \neq -3$, then $\text{rank}_S(T_g) = 4$.
- ▷ if $\lambda = -3$, then $\text{rank}_S(T_g) = 5$.

This gives us that

- ▷ $d = (3, 4, 1), r = 3$, $\mathcal{M}_{d,3} \subsetneq \mathcal{V}_{d,3} = \text{Sym}_3(\mathbb{R}^3)$.
- ▷ $d' = (3, 5, 1), r = 3$, $\mathcal{M}_{d',3} = \text{Sym}_3(\mathbb{R}^3)$.

Question: For a 2-layer network architecture (d, r) such that $\mathcal{V}_{d,r} \subsetneq (\text{Sym}_r(\mathbb{R}^{n_1}))^{n_h}$, are there any other examples (other than $d = (n, m, 1),\ r = 2$) where $\mathcal{M}_{d,r} = \mathcal{V}_{d,r}$?

Example: $d = (2, 2, 2, 1)$ **and** $r = 2$

For the architecture $d = (2, 2, 2, 1)$ and $r = 2$, we have the following polynomial map

$$p_\theta(\mathbf{x}) = (W_3 \rho_2 W_2 \rho_2 W_1)\mathbf{x} = W_3 \rho_2 (W_2 \rho_2 W_1 \mathbf{x}) = W_3 \rho_2 \begin{pmatrix} b_{11} q_1^2 + b_{12} q_2^2 \\ b_{21} q_1^2 + b_{22} q_2^2 \end{pmatrix} =$$

$$= \begin{pmatrix} c_1 & c_2 \end{pmatrix} \begin{pmatrix} (b_{11} q_1^2 + b_{12} q_2^2)^2 \\ (b_{21} q_1^2 + b_{22} q_2^2)^2 \end{pmatrix} = c_1 (b_{11} q_1^2 + b_{12} q_2^2)^2 + c_2 (b_{21} q_1^2 + b_{22} q_2^2)^2.$$

So, the image of p_θ is given by a homogeneous polynomial of degree 4 in two variables that can be decomposed as

$$\alpha_1 q_1^4 + \alpha_2 q_1^2 q_2^2 + \alpha_3 q_2^4$$

for some α_i depending on a_{ij}, b_{pq}, and c_k.

Question: What can we say about decomposing real symmetric tensors $T \in \mathrm{Sym}_4(\mathbb{R}^2)$ as $T = \alpha_1 v_1^{\otimes 4} + \alpha_2 v_1^{\otimes 2} v_2^{\otimes 2} + \alpha_2 v_2^{\otimes 4}$?

Thank you! Questions? Comments?