Activation thresholds and expressiveness of polynomial neural networks

Thomas Yahl
tyahl@wisc.edu
University of Wisconsin – Madison

JMM 2025

January 2025

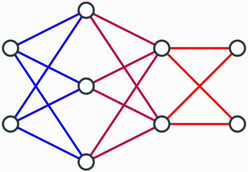
Joint work with Bella Finkel, Jose Israel Rodriguez, and Chenxi Wu

Neural Networks

- A (feedforward) neural network $F_{\boldsymbol{\theta}} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ is a composition of linear maps $A_i : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ and non-linear maps $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$,

$$F_{\boldsymbol{\theta}}(x) = (A_L \circ \sigma_{L-1} \circ A_{L-1} \circ \cdots \circ A_2 \circ \sigma_1 \circ A_1)(x).$$

  - The non-linear maps $\sigma_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ are coordinate-wise applications of a fixed function called the activation function.
  - A neural network is parameterized by $\boldsymbol{\theta} = (A_1, \ldots, A_L)$.
- The architecture of the neural network $F_{\boldsymbol{\theta}}$ is the sequence $\boldsymbol{d} = (d_0, \ldots, d_L)$.



Neural Networks

Common activation functions in applications are:

- Rectified Linear Unit: $R(x) = \max\{x, 0\}$
- Sigmoid function: $S(x) = \dfrac{e^x}{e^x + 1}$
- Gaussian Error Linear Unit:
  $G(x) = \dfrac{x}{2}\left[1 + \tanh\left(\sqrt{\dfrac{2}{\pi}}(x + 0.044715x^3)\right)\right]$

We will work with polynomial neural networks, whose activation function are given by power functions, $\sigma(x) = x^r$.

- The degree r is called the activation degree.

Polynomial Neural Networks

Given an architecture $\boldsymbol{d} = (d_0, \ldots, d_L)$ and activation function $\sigma(x) = x^r$, the parameter map

$$\Psi_{\boldsymbol{d},r} : \mathbb{R}^{d_L \times d_{L-1}} \times \cdots \times \mathbb{R}^{d_1 \times d_0} \to (\mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0}))^{d_L}$$

is defined on the input $\boldsymbol{\theta} = (A_1, \ldots, A_L)$ by $\Psi_{\boldsymbol{d},r}(\boldsymbol{\theta}) = F_{\boldsymbol{\theta}}$.

Definition
The neurovariety $\mathcal{V}_{\boldsymbol{d},r}$ is the Zariski closure of the image of $\Psi_{\boldsymbol{d},r}$.

- The neurovariety $\mathcal{V}_{\boldsymbol{d},r}$ is the closure of the set of functions that are representable as a neural network $F_{\boldsymbol{\theta}}$ with architecture $\boldsymbol{d}$ and activation function $\sigma(x) = x^r$.

Polynomial Neural Networks

Example: Consider the architecture $\boldsymbol{d} = (2, 2, 3)$ and activation function $\sigma(x) = x^2$. In this setting, writing $\boldsymbol{\theta} = (A, B)$, a polynomial neural network $F_{\boldsymbol{\theta}}$ has the form

$$F_{\boldsymbol{\theta}}(x, y) = B\sigma A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b_{11}(a_{11}x + a_{12}y)^2 + b_{12}(a_{21}x + a_{22}y)^2 \\ b_{21}(a_{11}x + a_{12}y)^2 + b_{22}(a_{21}x + a_{22}y)^2 \\ b_{31}(a_{11}x + a_{12}y)^2 + b_{32}(a_{21}x + a_{22}y)^2 \end{pmatrix}.$$

Thus, $\Psi_{\boldsymbol{d},r} : \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 3} \to (\mathrm{Sym}_2(\mathbb{R}^2))^3 = \mathbb{R}^9$ is defined by

$$\Psi_{\boldsymbol{d},r}(\theta) = (b_{11}a_{11}^2 + b_{12}a_{21}^2, b_{11}a_{11}a_{12} + b_{12}a_{21}a_{22}, b_{11}a_{12}^2 + b_{12}a_{22}^2, \dots).$$

The corresponding neurovariety $\mathcal{V}_{(2,2,3),2} \subseteq (\mathrm{Sym}_2(\mathbb{R}^2))^3 = \mathbb{R}^9$ is a hypersurface defined by

$$z_3z_5z_7 - z_2z_6z_7 - z_3z_4z_8 + z_1z_6z_8 + z_2z_4z_9 - z_1z_5z_9 = 0.$$

Polynomial Neural Networks

- The dimension $\dim \mathcal{V}_{\boldsymbol{d},r}$ is a meausure of the expressivity.
- From a parameter count, there is an expected dimension

$$\operatorname{edim} \mathcal{V}_{\boldsymbol{d},r} = \min \left\{ d_L + \sum_{i=0}^{L-1} d_{i+1}(d_i - 1), d_L \binom{d_0 + r^{L-1} - 1}{r^{L-1}} \right\}.$$

$\boldsymbol{d}$ \r	2	3	4	5	6
(2,2,2)	6	6	6	6	6
(2,3,2)	6	8	9	9	9
(4,5,3)	29	30	30	30	30
(3,5,6)	35	40	40	40	40
(2,3,2,2)	8	10	11	11	11

Figure 1: Table of dimensions of small neurovarieties

Polynomial Neural Networks

Theorem (Alexander,Hirschowitz)
If $\boldsymbol{d} = (d_0, d_1, 1)$, *then* $\dim \mathcal{V}_{\boldsymbol{d},r} = \operatorname{edim} \mathcal{V}_{\boldsymbol{d},r}$ *except in the following cases:*

- $r = 2$, $2 \leq d_1 \leq d_0 - 1$
- $r = 3$, $d_0 = 5$, $d_1 = 7$
- $r = 4$, $d_0 = 3$, $d_1 = 5$
- $r = 4$, $d_0 = 4$, $d_1 = 9$
- $r = 4$, $d_0 = 5$, $d_1 = 15$

Conjecture (Kileel,Trager,Bruna)
For all architectures $\boldsymbol{d}$, there exists $\widetilde{r} = \widetilde{r}(\boldsymbol{d})$ such that for $r > \widetilde{r}$,

$$\dim \mathcal{V}_{\boldsymbol{d},r} = \operatorname{edim} \mathcal{V}_{\boldsymbol{d},r}.$$

Activation Thresholds

Definition
The activation threshold $\operatorname{ActThr}(\boldsymbol{d})$ of an architecture $\boldsymbol{d}$, if it exists, is the smallest number $\widetilde{r} = \operatorname{ActThr}(\boldsymbol{d})$ such that if $r > \widetilde{r}$, then

$$\dim \mathcal{V}_{\boldsymbol{d},r} = \operatorname{edim} \mathcal{V}_{\boldsymbol{d},r}.$$

Theorem (Finkel,Rodriguez,Wu,Y.)
For all architectures $\boldsymbol{d}$*, the activation threshold* $\operatorname{ActThr}(\boldsymbol{d})$ *exists and is bounded above by*

$$\operatorname{ActThr}(\boldsymbol{d}) \leq 8\,(2\max \boldsymbol{d} - 1)^2 - 1.$$

- Our bounds on the activation threshold are derived from results on Waring's problem in number theory.

Activation Thresholds

Example: Consider the architecture $\boldsymbol{d} = (3, 2, 3, 2)$. The dimension and expected dimension of the neurovariety $\mathcal{V}_{\boldsymbol{d},r}$ are computed below for various values of r.

r	2	3	4	5	6
dim	10	12	13	13	13
edim	13	13	13	13	13

- The activation degree is bounded above by $\mathrm{ActThr}(d) \leq 199$, but it appears that the dimensions stabilize for $r > 3$.

Activation Thresholds

- An architecture $\boldsymbol{d} = (d_0, \ldots, d_L)$ is equi-width if $d_0 = d_1 = \cdots = d_L$ and non-increasing if $d_0 \geq d_1 \geq \cdots \geq d_L$.

Theorem
If $\boldsymbol{d} = (d_0, \ldots, d_L)$ is an equi-width architecture with $d_L > 1$, then $\operatorname{ActThr}(\boldsymbol{d}) = 1$*. That is, if $r > 1$, then* $\dim \mathcal{V}_{\boldsymbol{d},r} = \operatorname{edim} \mathcal{V}_{\boldsymbol{d},r}$.

Conjecture (Kubjas,Li,Wiesmann)
If $\boldsymbol{d} = (d_0, \ldots, d_L)$ is a non-increasing architecture with $d_L > 1$, then $\operatorname{ActThr}(\boldsymbol{d}) = 1$. That is, if $r > 1$, then $\dim \mathcal{V}_{\boldsymbol{d},r} = \operatorname{edim} \mathcal{V}_{\boldsymbol{d},r}$.

Activation Thresholds

Open Problems!

- How to compute better bounds for the activation threshold of an architecture?
- How to compute the exact activation threshold of an architecture?
- Are there generalizations of activation threshold to other activation functions that depend on parameters?

References i

Thank you all for your time!

M. L. Green.
Some Picard theorems for holomorphic maps to algebraic varieties.
Amer. J. Math., 97:43–75, 1975.

J. Kileel, M. Trager, and J. Bruna.
On the expressive power of deep polynomial neural networks.
Advances in neural information processing systems, 32, 2019.

K. Kubjas, J. Li, and M. Wiesmann.
Geometry of polynomial neural networks.
Algebr. Stat., 15(2):295–328, 2024.

D. Newman and M. Slater.
Waring's problem for the ring of polynomials.
Journal of Number Theory, 11(4):477–487, 1979.

B. Reznick.
Patterns of dependence among powers of polynomials, 2001.
arXiv:0106060.

Neural Networks

There are several well celebrated universal approximation theorems:

Theorem
A continuous function $\sigma : \mathbb{R} \to \mathbb{R}$ is not a polynomial if and only if for every continuous function $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_2}$, compact set C, and $\epsilon > 0$, there exists $d_1 > 1$ and a neural network $F_{\boldsymbol{\theta}}$ with architecture $\boldsymbol{d} = (d_0, d_1, d_2)$ and activation function σ such that

$$\sup_{x \in C} ||f(x) - F_{\boldsymbol{\theta}}(x)|| < \epsilon.$$

- So why bother considering polynomial neural networks?
  - It may not be possible to a priori compute a sufficient width d_1.
  - Not every continuous function may need to be approximated for a given application–more specific tools have more specific uses.