On the Geometry and Optimization of Polynomial Convolutional Networks

Vahid Shahverdi* **Giovanni Luca Marchetti*** **Kathlén Kohn***
KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

We study convolutional neural networks with monomial activation functions. Specifically, we prove that their parameterization map is regular and is an isomorphism almost everywhere, up to rescaling the filters. By leveraging on tools from algebraic geometry, we explore the geometric properties of the image in function space of this map – typically referred to as neuromanifold. In particular, we compute the dimension and the degree of the neuromanifold, which measure the expressivity of the model, and describe its singularities. Moreover, for a generic large dataset, we derive an explicit formula that quantifies the number of critical points arising in the optimization of a regression loss.

1 INTRODUCTION

Deep neural networks – and, in general, parametric machine learning models – define a space of functions as their parameters vary. These spaces are often referred to as *neuromanifolds* (Kohn, 2024; Calin, 2020). Understanding the geometry of neuromanifolds is a subtle yet fundamental challenge due to its intimate connection to the training process. Namely, neural networks learn by following a gradient flow that attracts the model to (an estimate of) the ground-truth function, which can be interpreted as minimizing a functional distance over the neuromanifold. Therefore, geometric problems over neuromanifolds – such as nearest point problems – are related to the learning dynamics of the corresponding models.

For neural networks with polynomial activation functions, the neuromanifold is a (semi-) algebraic set, i.e., it is defined by polynomial equalities and inequalities.

*Equal contribution.

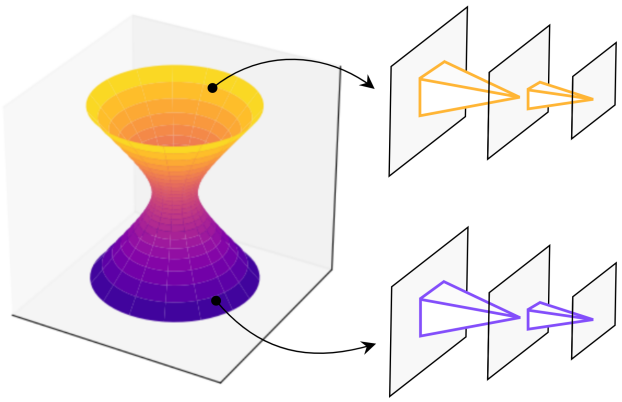


Figure 1: Illustration of a Segre–Veronese variety parametrizing CNNs.

This enables to exploit tools from the rich field of *algebraic geometry* for analyzing neuromanifolds. In particular, various algebro-geometric invariants can provide insights into fundamental machine learning aspects. For example, the *dimension* of the neuromanifold plays a central role in statistical learning theory[1] since, according to the Fundamental Theorem of Learning (Shalev-Shwartz and Ben-David, 2014), it controls the sample complexity of learnability. Moreover, the *degree* is a curvature-based invariant that controls the concentration of measure around the neuromanifold and, in particular, the approximation error of the given model (Basu and Lerario, 2023). Dimension and degree are, in other words, basic measures of expressivity. Lastly, the *Euclidean distance degree* (Draisma et al., 2016) is a modern invariant counting the singularities of the distance function over the neuromanifold from an external point. In particular, it upper-bounds the number of (locally-) nearest points. Since these invariants are well-understood for a wide class of algebraic varieties, they can provide insights into the learning process of networks of various architectures.

[1]In this context, neuromanifolds are referred to as 'hypothesis spaces', and are usually considered in a combinatorial version.

In this work, we consider deep *Convolutional* Neural Networks (CNNs) (Fukushima, 1979; LeCun et al., 1995) with monomial activation functions, and study their neuromanifolds. Our motivation is twofold. First, CNNs are popular models deployed in various signal processing domains. Historically, they have played a central role in several modern breakthroughs across deep learning (Krizhevsky et al., 2012; Van Den Oord et al., 2016), and have recently seen a resurgence in computer vision (Liu et al., 2022). Second, the convolutional architecture is particularly suitable for algebraic geometry. Indeed, the convolution is a well-behaved algebraic operation, which translates into geometric properties for the associated neuromanifold and its parametrization.

1.1 Summary of Results

We provide theoretical results on both the geometry and the optimization of polynomial CNNs. Our arguments involve tools and ideas from algebraic geometry. In order to analyze neuromanifolds of polynomial CNNs, we study their parametrization. In this context, our main result states, informally, the following.

Informal Theorem 1.1 (Section 4.1)**.** *Up to rescaling each filter, the parameterization of the neuromanifold of a polynomial CNN is regular and is an isomorphism almost everywhere. The remaining fibers are finite.*

Intuitively, this shows that CNNs are parametrized in an 'optimal way' since, once the scaling symmetries are factored out, the parameters are in a one-to-one regular correspondence (almost everywhere) with the associated functions. This is in contrast with other neural architectures – such as fully-connected networks – where the parametrization exhibits large fibers and critical points, translating into redundancy in the parameter space (Kileel et al., 2019).

An immediate consequence of Informal Theorem 1.1 is an expression for the dimension and the degree of the neuromanifold – see Corollary 4.7. In particular, we observe that the dimension grows linearly w.r.t. the number of layers, while the degree grows (super-) exponentially. This provides an intuitive explanation to the importance of depth in neural networks; neuromanifolds of deeper CNNs remain low-dimensional – translating into a low sample complexity of learning – while efficiently filling their ambient function space – translating into expressive power.

Informal Theorem 1.1 implies other geometric properties. First, it follows that the neuromanifold is closed in the Zariski topology of its ambient space (see Proposition 4.8), and therefore in the Euclidean one as well. In terms of learning dynamics, this shows that the neuromanifold contains its asymptotic limits. Another consequence is the description of *singularities* of the neuromanifold, i.e., special CNNs where the neuromanifold looks degenerate. It follows from Informal Theorem 1.1 that the singular points are the simplest possible – specifically 'nodal' singularities – which, as discussed below, translates into advantages for the optimization dynamics. Such singularities can only come from 'subnetworks', meaning that they arise when (specific) weights vanish.

In order to prove Informal Theorem 1.1, we first remove the scaling symmetries by working in the projective space, and then to rely on tools from projective geometry. In particular, we leverage on birational geometry which, roughly speaking, concerns with properties that are preserved almost everywhere. Moreover, we factorize the projectified parametrization into a linear projection and the *Segre–Veronese* embedding. The latter is a classical map in algebraic geometry whose properties are well-understood – see Figure 1 for an illustration.

As anticipated before, understanding the geometry of the neuromanifold enables us to study the learning dynamics of polynomial CNNs. In particular, we consider the square-error regression loss, and prove the following.

Informal Theorem 1.2 (Section 4.2)**.** *For large generic datasets, the square-error loss can be rephrased as a distance minimization over the neuromanifold from an external polynomial function. There is an explicit formula for the number of (complex) critical points of the loss.*

The formula above is obtained via the theory of the Euclidean distance degree, which has a known expression for the Segre–Veronese variety. Surprisingly, for almost all datasets, the number of critical points over the neuromanifold is finite and does not depend on the dataset nor on its size (assuming it is large enough). However, the theory requires taking complex solutions into account. Therefore, our formula provides an exact count for complex critical points, as well as an upper bound for the real ones. Since local minima are critical points, our formula bounds their number as well. Finally, we argue that counting critical points over the neuromanifold is (almost) equivalent to counting them, up to scaling, over the parameter space, which is where optimization is performed in practice. This involves showing that the singular points of the neuromanifold are not critical for the distance function (with one exception), which is possible via our description of singularities.

2 RELATED WORK

Since our work is concerned with neuromanifolds of polynomial CNNs, we review the literature around neuromanifolds and polynomial neural networks.

Algebraic Geometry of Neuromanifolds. Neuromanifolds have been analyzed via algebraic geometry in several instances. Fully-connected polynomial networks have been discussed in (Kileel et al., 2019; Finkel et al., 2024), with a focus on problems regarding dimensionality, while linear CNNs have been discussed in (Kohn et al., 2022, 2023; Shahverdi, 2024), with a focus on singularities and critical points of the loss function. More recently, neuromanifolds of linear self-attention networks have been considered (Henry et al., 2024). Here, we contribute to this line of research by discussing the geometry of neuromanifolds defined by polynomial CNNs. While the previous literature focuses mainly on the linear case – with only partial results for the general polynomial one – our work is the first to provide a comprehensive description of both the parametrization and the neuromanifold of a polynomial model.

Polynomial Activation Functions. Standard activation functions for neural networks – such as the Rectified Linear Unit (ReLU) – are not polynomial. Nonetheless, networks with polynomial activations have been considered in the literature. Such networks are universal interpolators and approximators (Constantinescu and Popescu, 2023). Various flavors of polynomial activations have been discussed, ranging from monomial versions of ReLU (Berradi, 2018; Li et al., 2019), to rational functions (Boullé et al., 2020; Telgarsky, 2017), to piece-wise polynomial functions (López-Rubio et al., 2019; Hou et al., 2017). Finally, quadratic activations have appeared in neuroscience for modelling biological neural networks (Adelson and Bergen, 1985).

3 BACKGROUND

In this section, we introduce the necessary background around convolutional neural networks, as well as the relevant tools from algebraic geometry.

3.1 Polynomial Convolutional Networks

We start by introducing convolutional neural networks. For simplicity of notation, we focus on one-dimensional convolutions – all theory and results can be extended verbatim to the higher-dimensional case. Given integers $k, s, d' \in \mathbb{N}$ representing filter size, stride, and output dimension respectively, the convolution between a filter $w \in \mathbb{R}^k$ and an input vector $x \in \mathbb{R}^d$, with $d = s(d'-1)+k$, is the vector $w \star_s x \in \mathbb{R}^{d'}$ defined for $0 \le i < d'$ as:

$$(w \star_s x)[i] = \sum_{0 \le j < k} w[j]\ x[si + j]. \tag{1}$$

The convolution in Equation (1) is linear in x and is represented by a $d' \times d$ *Toeplitz matrix*. For example, given $k = 3$, $s = 2$, and $d' = 3$, the corresponding matrix is

$$\begin{pmatrix} w[0] & w[1] & w[2] & 0 & 0 & 0 & 0 \\ 0 & 0 & w[0] & w[1] & w[2] & 0 & 0 \\ 0 & 0 & 0 & 0 & w[0] & w[1] & w[2] \end{pmatrix}.$$

Moreover, the composition of convolutions can be rephrased as polynomial multiplications. We associate to the filter the following homogeneous bivariate polynomial of degree $k-1$ in (a^s, b^s):

$$\pi_s(w) = \sum_{0 \le i < k} w[i]\ a^{s(k-i-1)} b^{si}. \tag{2}$$

Then the following property holds: an iterated convolution $v \star_s (w \star_t x)$ with respective strides s, t coincides with a convolution $q \star_{st} x$ whose associated polynomial satisfies $\pi_1(q) = \pi_t(v)\ \pi_1(w)$.

Convolutions can be composed in order to construct deep convolutional networks. To this end, fix an integer $L \in \mathbb{N}$ and sequences $\mathbf{k}, \mathbf{s} \in \mathbb{N}^L$, $\mathbf{d} \in \mathbb{N}^{L+1}$ such that $d_i = s_i(d_{i+1} + k_i)$ for all i. Moreover, consider a map $\sigma\colon \mathbb{R} \to \mathbb{R}$ playing the role of the activation function.

Definition 1. A *Convolutional Neural Network* (CNN) with weights $\mathbf{w} = (w_0, \dots, w_{L-1}) \in \bigoplus_{0 \le i < L} \mathbb{R}^{k_i}$ is the map $\varphi_{\mathbf{w}}\colon \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ given by:

$$\varphi_{\mathbf{w}}(x) = w_{L-1} \star_{s_{L-1}} \sigma\left(\cdots \star_{s_1} \sigma(w_0 \star_{s_0} x)\right), \tag{3}$$

where σ is applied coordinate-wise.

Given $r \in \mathbb{N}$, denote by $\sigma_r(x) = x^r$ the power function. CNNs with activation function $\sigma = \sigma_r$ are called *polynomial*, and they are called *linear* if $r = 1$ i.e., if σ is the identity function. Polynomial CNNs define homogeneous polynomial functions of degree r^{L-1}, meaning that $\varphi_{\mathbf{w}} \in \mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})^{d_L}$. Here, $\mathrm{Sym}_\alpha(\mathbb{R}^\beta)$ denotes the space of symmetric tensors of degree α in β variables, i.e., the vector space of dimension $\binom{\beta+\alpha-1}{\alpha}$ whose canonical basis is given by monomials of degree α in $x[0], \dots, x[\beta-1]$.

Definition 2. The *neuromanifold* of a polynomial CNN is the image in $\mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})^{d_L}$ of the parametrization, i.e.,

$$\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r} = \left\{ \varphi_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^{|\mathbf{k}|} \right\}, \tag{4}$$

where $|\mathbf{k}| = k_0 + \cdots + k_{L-1}$. By the Tarski-Seidenberg theorem, the neuromanifold is a semi-algebraic set, i.e., it can be defined by polynomial equalities and inequalities. However, we will show that it is actually an algebraic *variety*, meaning that it is closed in Zariski topology of $\mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})^{d_L}$, or, equivalently, that it can be defined by polynomial equalities alone – see Proposition 4.8. Lastly, throughout this work we will often consider complex coefficients. To this end, we denote by $\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ the complexification of the neuromanifold. The latter can be defined by replacing real numbers $\mathbb{R}$ with complex ones $\mathbb{C}$ in all definitions of this section.

3.2 Segre–Veronese Varieties

Here, we recall Segre–Veronese varieties and their basic properties. As we will show, these are closely related to the neuromanifolds of CNNs. To this end, fix $k \in \mathbb{N}$, and let $\mathbf{m}, \mathbf{p} \in \mathbb{N}^k$. Consider k vector spaces $V_1, \ldots, V_k$ with $\dim(V_i) = p_i + 1$ over a field $\mathbb{K}$. Denote by $\mathbb{P}V = V/(\mathbb{K} \setminus \{0\})$ the projectification of a vector space V.

Definition 3. The *Segre–Veronese embedding* is the map

$$\nu_{\mathbf{m},\mathbf{p}} \colon \prod_{1 \le i \le k} \mathbb{P}V_i \to \mathbb{P} \bigotimes_{1 \le i \le k} \mathrm{Sym}_{m_i}(V_i) \qquad (5)$$

defined by taking tensor products of symmetric powers of vectors in the corresponding spaces. The *Segre–Veronese variety* $\mathcal{V}_{\mathbf{m},\mathbf{p}}$ is the image of $\nu_{\mathbf{m},\mathbf{p}}$.

Explicitly, if $V_i = \mathbb{R}^{p_i+1}$, then the Segre–Veronese variety consists of (tensor) products of k monomials of corresponding degree m_i. In particular, $\nu_{\mathbf{m},\mathbf{p}}$ is an embedding, as suggested by the nomenclature, and $\mathcal{V}_{\mathbf{m},\mathbf{p}}$ is a smooth projective variety of dimension $|\mathbf{p}| = p_1 + \cdots + p_k$. When $k = 1$, we refer to $\nu_{m,p}$ simply as Veronese embedding. For example, for $k = 2$ and $p_1 = p_2 = m_1 = m_2 = 1$, $\mathcal{V}_{\mathbf{m},\mathbf{p}}$ coincides with a smooth quadric in the three-dimensional projective space $\mathbb{P}(V_1 \otimes V_2)$, which can be represented as a hyperboloid – see Figure 1.

3.3 Euclidean Distance Degree

We now recall the *Euclidean distance degree* (Draisma et al., 2016) – an invariant in algebraic geometry that will be central in our work. Intuitively, this invariant counts the number of critical points of the Euclidean distance function from (the smooth locus of) an algebraic variety to a fixed external point. Formally, let $X \subset \mathbb{R}^n$ be an algebraic variety, whose smooth locus is denoted by X_{reg}. Let A be a $n \times n$ symmetric positive-definite matrix inducing a distance $\mathrm{d}_A(x,y)^2 = (x-y)^\top A\,(x-y)$ for $x, y \in \mathbb{R}^n$. We consider the (squared) distance function from an anchor $u \in \mathbb{R}^n \setminus X$ to the smooth locus of X:

$$\begin{array}{ccc} X_{\mathrm{reg}} & \to & \mathbb{R}_{\ge 0} \\ x & \mapsto & \mathrm{d}_A(x,u)^2. \end{array} \qquad (6)$$

For what follows, it is necessary to consider complex coefficients. To this end, we complexify the above map, extending it as $X^{\mathbb{C}}_{\mathrm{reg}} \to \mathbb{C}$, where $X^{\mathbb{C}} \subset \mathbb{C}^n$ denotes the complexification of X.

Definition 4. The *Euclidean distance degree* of X with respect to A is the number of complex critical points of the complexified map $X^{\mathbb{C}}_{\mathrm{reg}} \to \mathbb{C}$ for generic u.

Geometrically, a point $x \in X_{\mathrm{reg}}$ is critical for the distance map if, and only if, $x - u$ is perpendicular to the tangent space of X at x according to the scalar product induced by A. Moreover, it is possible to extend the notion of Euclidean distance degree to projective varieties $X \subset \mathbb{PR}^n$ by considering their de-projectification, i.e., the corresponding affine cone in $\mathbb{R}^n$.

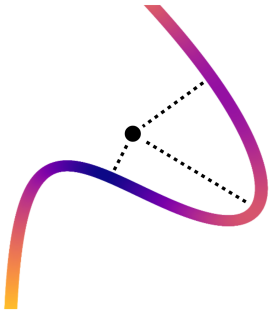


Figure 2: Distance function from an anchor to a curve, visualized as a color gradient. The critical values are denoted by dotted lines.

For almost all A, the Euclidean distance degree is the same (Draisma et al., 2016, Theorem 6.11). This number is referred to as *generic* Euclidean distance degree of X, denoted by $\mathrm{gED}(X) \in \mathbb{N}$. The following result – which will be necessary in our forthcoming discussion – provides a closed formula for the generic Euclidean distance degree of Segre–Veronese varieties.

Theorem 3.1 (Kozhasov et al. (2023))**.** *The generic Euclidean Distance degree of the Segre–Veronese variety is:*

$$\mathrm{gED}(\mathcal{V}_{\mathbf{m},\mathbf{p}}) = \sum_{0 \le i \le |\mathbf{p}|} (-1)^i \left(2^{|\mathbf{p}|+1-i} - 1\right) (|\mathbf{p}| - i)! \sum_{\substack{|\alpha| = i \\ \forall j\ \alpha_j \le p_j}} \prod_{1 \le j \le k} \frac{\binom{p_j+1}{\alpha_j}}{(p_j - \alpha_j)!}\, m_j^{p_j - \alpha_j},$$

where $|\mathbf{p}| = p_1 + \cdots + p_k$.

4 CONVOLUTIONAL NEUROMANIFOLDS

In this section, we study the neuromanifolds of polynomial CNNs, and present the main results of this work. We first discuss geometric properties – such as dimension, projectification, and relation to Segre–Veronese varieties – and then proceed to compute the Euclidean distance degree of the neuromanifold. We provide most of the proofs in the appendix – see Section B.

4.1 Geometry

Throughout this section, we will leverage on principles and tools from algebraic geometry. Therefore, it will be convenient to work with complex coefficients. For simplicity, we will abuse the notation and use the symbols from Section 3.1 to denote the analogous complex object, obtained by replacing $\mathbb{R}$ with $\mathbb{C}$ in the corresponding definition. While all the results of this section are stated and proved over complex numbers, almost all of them extend a posteriori to the real case – see Remark 4.3.

We start with a simple but powerful property of convolutions.

Lemma 4.1. *For $w \in \mathbb{C}^k \setminus \{0\}$ and $s \in \mathbb{N}$, the convolution map $x \mapsto w \star_s x$ has full rank.*

Proof. See Section B.1 in the appendix. □

Consider now a polynomial CNN $\varphi_{\mathbf{w}}$ parametrized by weights $\mathbf{w} \in \mathbb{C}^{|\mathbf{k}|}$.

Corollary 4.2. *If $\varphi_{\mathbf{w}}(x) = 0$ for all $x \in \mathbb{C}^{d_0}$, then $w_i = 0$ for some i.*

Proof. Suppose that $w_i \neq 0$ for all i. By Lemma 4.1, for all i the map $x \mapsto \sigma_r(w_i \star_{s_i} x)$ is open, i.e., it sends open sets to open ones. Openness is preserved by composition and, in particular, $\varphi_{\mathbf{w}} \neq 0$, as desired. □

Remark 4.1. The property established by Corollary 4.2 is characteristic of (polynomial) CNNs, and does not hold for other network architectures. For example, a fully-connected network with arbitrary activations and at least two layers will not satisfy it. To illustrate this, consider a 2-layer fully-connected network $W_1 \cdot \sigma(W_0 \cdot x)$, where $x \in \mathbb{R}^{d_0}$ is the input, $W_0 \in \mathbb{R}^{d_1 \times d_0}$ and $W_1 \in \mathbb{R}^{d_2 \times d_1}$ are the weight matrices, and $\sigma\colon \mathbb{R} \to \mathbb{R}$ is the activation function. Assuming $d_1 > 1$, if the first two rows of W_0 coincide and the other ones vanish, while the first two columns of W_1 are opposite, then the network vanishes on all $x \in \mathbb{R}^{d_0}$.

As a consequence of Corollary 4.2, it is possible to projectify the neuromanifold and its parametrization. Indeed, the result implies that the map $\mathbf{w} \mapsto \varphi_{\mathbf{w}}$ parametrizing the neuromanifold induces an algebraic morphism between the corresponding projective spaces:

$$\overline{\varphi}\colon \prod_{0 \leq i < L} \mathbb{P}\mathbb{C}^{k_i} \to \mathbb{P}\mathrm{Sym}_{r^{L-1}}(\mathbb{C}^{d_0})^{d_L}. \tag{7}$$

We denote by $\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ the image of $\overline{\varphi}$ and deem it *projective neuromanifold*. We deduce the following relevant property of the (projective) neuromanifold.

Proposition 4.3. *Both $\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ and $\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ are closed in the Zariski topology of their respective ambient spaces.*

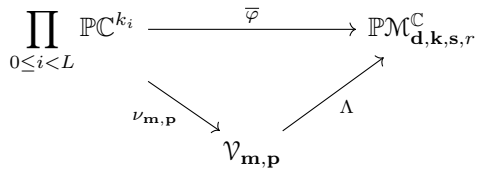
Proof. See Section B.2 in the appendix. □

In order to analyse $\overline{\varphi}$, we provide another technical result that enables to rephrase polynomial CNN layers in terms of convolutions. To this end, fix a filter $w \in \mathbb{C}^k$, a stride s, and an exponent r.

Proposition 4.4. *For every $x \in \mathbb{C}^n$, $\sigma_r(w \star_s x)$ can be written as a convolution $\tilde{w} \star_{\tilde{s}} \tilde{x}$ with stride $\tilde{s} = s\binom{r+k-1}{r}$ and filter size $\tilde{k} = \binom{r+k-1}{r}$. Moreover, $\tilde{x}$ and $\tilde{w}$ consist of monomials in x and w, respectively.*

Proof. See Section B.3 in the appendix. □

Proposition 4.4 implies, inductively, that $\overline{\varphi}$ factors via the Segre–Veronese embedding (Definition 3). Specifically, let $\mathbf{m} = (r^{L-1}, \dots, r, 1)$ and $\mathbf{p} = (k_0 - 1, \dots, k_{L-1} - 1)$. Then the following diagram commutes:



Here, Λ is (the projectification of) a linear map restricted to the Segre–Veronese variety. In the next section, this factorization enables us to compute the neuromanifold's generic Euclidean distance degree.

We now provide the main results of this section. Recall that a differentiable map is *regular* if its differential has maximal rank at every point of the domain.

Theorem 4.5. *Assume that $r > 1$. Then the projectified parametrization $\overline{\varphi}$ is regular.*

Proof. The argument requires several technical computations around the differential of φ, whose details

$r = 2$ $r = 5$ $r = 10$

Figure 3: Visualization of two-dimensional neuromanifolds (over $\mathbb{R}$) corresponding to $(k_0, k_1) = (2, 2)$ projected orthogonally to $\mathbb{R}^3$, with varying activation degree.

we provide in the appendix, Section A. By Euler's theorem on homogeneous functions, showing that the differential of $\overline{\varphi}$ is injective at some $\mathbf{w}$ is equivalent to showing that the kernel of the differential of φ at $\mathbf{w}$ has dimension $L-1$. The latter follows immediately from Proposition A.2, which provides $L-1$ independent generators for such kernel. □

Next, we show that the map $\overline{\varphi}$ is *birational*, meaning that it is generically one-to-one. In other words, 'almost all' the fibers of $\overline{\varphi}$ are singletons. We also show that the remaining non-singleton fibers are finite.

Theorem 4.6. *Assume that $r > 1$. If $\overline{\varphi}_{\mathbf{w}} = \overline{\varphi}_{\mathbf{v}}$ for some $\mathbf{w}, \mathbf{v} \in \prod_i \mathbb{PC}^{k_i}$, then $\mathbf{w}$ and $\mathbf{v}$ are related by a shift of their indices. More precisely, there exist $t_0, \dots, t_{L-1}$ such that for all i:*

$$w_i = (0, \dots, 0, v_i[0], \dots, v_i[k_i - t_i - 1]), \tag{8}$$
$$v_i = (v_i[0], \dots, v_i[k_i - t_i - 1], 0, \dots, 0), \tag{9}$$

or vice versa. In particular, $\overline{\varphi}$ is birational on its image, and all the fibers are finite.

Proof. The argument involves a delicate induction over the number of layers – see Section B.4 in the appendix. □

Theorem 4.6 can be intuitively interpreted as stating that the projective neuromanifold is isomorphic to a Segre–Veronese variety 'almost everywhere'. For $r = 1$ – i.e., for linear CNNs – the same result has been shown by Kohn et al. (2023) with the additional assumption on strides $s_i > 1$ for all $0 \le i \le L-2$. Moreover, in the same work the authors show that the parametrization of linear CNNs has critical points, in contrast to the polynomial case. The argument is different from ours and is based on the interpretation of convolutions as polynomial multiplication.

Remark 4.2. From the proof of Theorem 4.6, it is evident that the shifts must satisfy arithmetic constraints involving the strides. Specifically, consider the shifts t_i together with their sign, which is set positive if w_i has zeros on the left, and negative otherwise. Define recursively $\tilde{t}_{-1} = 0$, $\tilde{t}_i = t_i + \tilde{t}_{i-1}/s_{i-1}$ for $i \ge 0$. Then $\tilde{t}_i$ must be an integer, and moreover $\tilde{t}_{L-1}$ must vanish. By the definition of (iterated) convolutions, this implies that if we truncate each filter by removing its zeros on the right or on the left, the resulting CNN coincides with $\varphi_{\mathbf{w}}$, albeit on a restricted domain with less input variables. Put simply, the CNNs whose (projectified) fibers are not singletons are the ones that can be defined by smaller architectures, i.e., special fibers arise from 'subnetworks'.

An immediate consequence of Theorem 4.6 is an expression for the two fundamental invariants of the neuromanifold: the dimension and the degree. Recall that for a given projective/affine variety, the latter counts, roughly speaking, the number of intersections with a generic linear subspace of dimension equal to the codimension of the variety.

Corollary 4.7. *The dimension and degree of the neuromanifold are:*

$$\dim(\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) = |\mathbf{k}| - L + 1, \tag{10}$$
$$\deg(\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) = (|\mathbf{k}| - L)! \prod_{0 \le j < L} \frac{r^{(L-j-1)(k_j-1)}}{(k_j-1)!}. \tag{11}$$

Proof. See Section B.5 in the appendix. □

In particular, note that if all the filter sizes $k_i := k$ are equal, then the degree is $(L(k-1))!\, r^{L(L-1)(k-1)/2}/(k-1)!^L$, which grows faster than $r^{L^2/2}$ w.r.t. L. Therefore, as anticipated in Section 1.1, the dimension grows linearly w.r.t. the depth of the network, while the degree grows (super-) exponentially.

Another consequence of the main results in this section is a description of the singular points of the neu-

romanifold. Since $\overline{\varphi}$ is regular and finite, the singularities of $\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ coincide with CNNs whose (projectified) fiber is not a singleton. Such singularities are of *nodal* type, i.e., points where the neuromanifold self-intersects, creating a finite number of (potentially overlapping) tangent spaces – see Figure 3 (center and right) for an illustration. By Remark 4.2, singularities arise from 'subnetworks'. Since $\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ is the affine cone of its projectification, the same description of singularities holds for the non-projective neuromanifold, apart from the base of the cone $\varphi_{\mathbf{w}} = 0$, which is a non-nodal singularity.

Remark 4.3. As anticipated, we remark that most of the results from this section hold over $\mathbb{R}$ as well. This is the case for Corollary 4.2 and Proposition 4.4 – since they are purely algebraic – as well as for Corollary 4.7 – since the dimension is independent of the coefficients – and for Theorem 4.5 and Theorem 4.6 – since birationality and rank of differential are invariant with respect to restriction of coefficients. However, the proof of Proposition 4.3 does not hold over $\mathbb{R}$, since it leverages on properties requiring algebraic closedness. Yet, in the following, we show that the main results of this section imply that the neuromanifold is still closed in the Zariski topology over $\mathbb{R}$.

Proposition 4.8. *Assume that $r > 1$. Then both $\mathbb{P}\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ and $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ are closed in the (real) Zariski topology of their respective ambient spaces.*

Proof. See Section B.6 in the appendix. □

In particular, the (projective) neuromanifold is a real algebraic variety. This does not hold for linear CNNs ($r = 1$), in which case the closure in the Zariski topology adds points to the neuromanifold (Kohn et al., 2023).

4.2 Optimization

We now discuss aspects of optimization of a polynomial CNN for a regression task. In particular, by leveraging on the theory of the Euclidean distance degree introduced in Section 3.3, we compute the number of (complex) critical points over the neuromanifold for the regression objective.

We start by introducing the regression objective. Consider a dataset, i.e., a finite subset $\mathcal{D} \subset \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ representing input-output pairs. The square-error loss is given by:

$$\mathcal{L}_{\mathcal{D}}(\varphi_{\mathbf{w}}) = \sum_{(x,y)\in\mathcal{D}} \|\varphi_{\mathbf{w}}(x) - y\|^2. \tag{12}$$

Training a CNN on the dataset $\mathcal{D}$ amounts to minimizing $\mathcal{L}_{\mathcal{D}}$. Intuitively, $\varphi_{\mathbf{w}}$ is optimized to interpolate $\mathcal{D}$, which is reminiscent of a closest point problem over the neuromanifold, but for a discrete set $\mathcal{D}$ instead of an anchor function. In the following, we show that for large datasets, this problem actually coincides with a closest-point problem.

Theorem 4.9. *There exists a linear subspace $V \subset \mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})^{d_L}$ containing the neuromanifold with the following property. For $|\mathcal{D}| \gg 0$, there exists a positive-definite quadratic form $A_{\mathcal{D}}$ over V – inducing a distance $\mathrm{d}_{A_{\mathcal{D}}}$ – and $u_{\mathcal{D}} \in V$ such that $\mathcal{L}_{\mathcal{D}}$ and $\mathrm{d}_{A_{\mathcal{D}}}(\bullet, u_{\mathcal{D}})^2$ coincide up to an additive constant. Moreover, $A_{\mathcal{D}}$ and $u_{\mathcal{D}}$ are generic for generic $\mathcal{D}$ of fixed cardinality.*

Proof. Let X and Y be the matrices whose columns are, respectively, $\nu_{r^{L-1},d_0-1}(x) \in \mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})$ and $y \in \mathbb{R}^{d_L}$ for $(x, y) \in \mathcal{D}$ (in some fixed order), where ν is the Veronese embedding. It follows that X has full rank for a sufficiently large $|\mathcal{D}|$. Consider the positive definite symmetric matrix $A_{\mathcal{D}} = XX^\top \otimes I_{d_L}$. By Kohn et al. (2022, Section 6), we have

$$\mathcal{L}_{\mathcal{D}}(\varphi_{\mathbf{w}}) = \mathrm{d}_{A_{\mathcal{D}}}(\varphi_{\mathbf{w}}, v_{\mathcal{D}})^2 + \mathrm{const}, \tag{13}$$

where $v_{\mathcal{D}} = YX^\top A_{\mathcal{D}}^{-1}$. Note that for generic $\mathcal{D}$, the quadratic form induced by $A_{\mathcal{D}}$ is not generic for $d_L > 1$. However, by Proposition 4.4, we can write

$$\varphi_{\mathbf{w}}(x) = \tilde{w} \star_{\tilde{s}} \tilde{x}, \tag{14}$$

where $\tilde{x}$ consists of (possibly repeated) monomials in x. Let V the space of linear maps $\mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0}) \to \mathbb{R}^{d_L}$ that can be written in the form of Equation 14 for some filter $\tilde{w}$. Such linear maps are determined by one of their output coordinates, meaning that the projections $V \to \mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})$ are injective. Now, $A_{\mathcal{D}}$ contains as a tensor factor the quadratic form $XX^\top$, which is generic over $\mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})$. It follows that $A_{\mathcal{D}}$ is generic when restricted to V. Lastly, let $u_{\mathcal{D}}$ be the orthogonal projection of $v_{\mathcal{D}}$ to V w.r.t. the scalar product induced by $A_{\mathcal{D}}$. Replacing $v_{\mathcal{D}}$ by $u_{\mathcal{D}}$ in Equation 14 introduces an additive constant, which concludes the proof. □

Remark 4.4. Similarly to Remark 4.1, we point out that the above result is characteristic of CNNs, and does not hold for other architectures. From the proof of Theorem 4.9 it is evident that both $A_{\mathcal{D}}$ and $u_{\mathcal{D}}$ can be defined for arbitrary polynomials in $\mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})^{d_L}$. However, in general, $A_{\mathcal{D}}$ is not generic, since it contains a tensor factor equal to the identity on the output space $\mathbb{R}^{d_L}$. For CNNs, this is circumvented by exploiting the fact that the first output entry determines the other ones, which follows from the equivariance properties of convolutions. In general, this does not hold. For example, for fully-connected polynomial networks, $A_{\mathcal{D}}$ is generic only if

$d_L = 1$ – i.e., for scalar-valued networks – and it is non-generic otherwise (Kubjas et al., 2024). In the latter case, the theory of the generic Euclidean distance degree is, unfortunately, inapplicable.

Theorem 4.9 draws a connection between the training of polynomial CNNs and (generic) distance minimization. Motivated by this, we compute the generic Euclidean distance degree of the neuromanifold.

Proposition 4.10. *Assume that $r > 1$ and let $\overline{k} = \dim(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) - 1 = |\mathbf{k}| - L$. The generic Euclidean distance degree of the neuromanifold is given by:*

$$\mathrm{gED}\left(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}\right) = \sum_{0\le i\le \overline{k}} (-1)^i \left(2^{\overline{k}+1-i} - 1\right) (\overline{k}-i)! \sum_{\substack{|\alpha|=i \\ \forall j\ \alpha_j<k_j}} \prod_{0\le j<L} \frac{\binom{k_j}{\alpha_j}}{(k_j-\alpha_j-1)!}\, r^{(L-j-1)(k_j-\alpha_j-1)},$$

Proof. See Section B.7 in the appendix. □

From Theorem 4.9 it follows that the formula from Proposition 4.10 counts the number of critical points over $(\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r})_{\mathrm{reg}}$ of the (complexified) loss function $\mathcal{L}_{\mathcal{D}}$ for large generic datasets $\mathcal{D}$. Since real critical points over a variety remain such after complexification, our formula provides an upper bound for the number of critical points of the loss function over $(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r})_{\mathrm{reg}}$.

A subtle point is that, in practice, optimization is performed via gradient descent over the parameter space, and not over the neuromanifold. The gradient flow of the loss – and in particular the number of critical points – could in principle differ due to the parametrization. However, this is not the case for polynomial CNNs: since the projectified parametrization is regular (Theorem 4.5), $\mathbf{w}$ is critical for $\mathcal{L}_{\mathcal{D}} \circ \varphi$ if, and only if, $\varphi_{\mathbf{w}}$ is critical for $\mathcal{L}_{\mathcal{D}}$. Therefore, our formula equivalently counts the number of critical points of $\mathcal{L}_{\mathcal{D}} \circ \varphi$ over $\varphi^{-1}((\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r})_{\mathrm{reg}}) \subseteq \mathbb{C}^{|\mathbf{k}|}$, up to the fibers of the parametrization, i.e., up to rescaling each filter.

Lastly, note that so far we have removed the singular points from the neuromanifold in the above constructions. Therefore, we now discuss the role of singular points of the neuromanifold in the optimization. The following result shows that, excluding the trivial vanishing CNN, the (parameters of) such singular points are not critical for a generic distance minimization problem.

Proposition 4.11. *Assume that $r > 1$. Let $V \subseteq \mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})^{d_L}$ be a linear subspace containing the neuromanifold, and A be a positive-definite quadratic form on V. For a generic element $u \in V$, if $\mathbf{w} \in \mathbb{R}^{|\mathbf{k}|}$ is critical for $\mathrm{d}_A(\varphi_\bullet, u)^2$, then either $\varphi_{\mathbf{w}} = 0$ or $\varphi_{\mathbf{w}} \in (\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r})_{\mathrm{reg}}$.*

Proof. See Section B.8 in the appendix. □

Therefore, if we exclude vanishing filters, we obtain the number of critical points in parameter space, up to rescaling each filter, of $\mathcal{L}_{\mathcal{D}} \circ \varphi$ for large generic $\mathcal{D}$.

5 CONCLUSIONS AND FUTURE WORK

In this work, we have studied aspects of the geometry and optimization of polynomial CNNs. In particular, we have proven that the projectified parametrization is regular and finite birational, derived the dimension of the neuromanifold, and related the latter to the Segre–Veronese variety. Moreover, we have rephrased the optimization of the regression loss as a distance minimization problem, and leveraged on the theory of the Euclidean distance degree to compute the number of (complex) critical points for a generic large dataset.

Our tools involve general yet powerful results from algebraic geometry. Therefore, similar arguments might be applicable to other neural architectures. For example, *graph neural networks* (Kipf and Welling, 2016; Bronstein et al., 2021) are nowadays popular in a variety of domains, and are closely related to CNNs. Therefore, exploring applications of algebraic geometry to the neuromanifold of such networks represents an interesting line for future investigation. From a broader perspective, a fundamental challenge lies in extending the study of neuromanifolds beyond the algebraic setting, i.e., for activation functions that are not polynomial (e.g., sigmoid and ReLU).

A limitation of the Euclidean distance degree is that it quantifies the number of critical points, without distinguishing local maxima, minima, or saddle points. Since local minima are the stable equilibria of gradient descent to which the latter converges, counting them represents a fundamental challenge from the perspective of optimization and learning. To this end, tools from *Morse theory* (Milnor, 1963) – an approach relating critical points with prescribed index to topological invariants – might be applicable to neuromanifolds, providing insights into local minima of the loss function. This constitutes another interesting direction to investigate.

Acknowledgements

We thank Rainer Sinn for helpful discussions on real algebraic geometry and the recovery of polynomials

from partial terms. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299.

Basu, S. and Lerario, A. (2023). Hausdorff approximations and volume of tubes of singular algebraic sets. *Mathematische Annalen*, 387(1-2):79–109.

Berradi, Y. (2018). Symmetric power activation functions for deep neural networks. In *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications*, pages 1–6.

Boullé, N., Nakatsukasa, Y., and Townsend, A. (2020). Rational neural networks. *Advances in neural information processing systems*, 33:14243–14253.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.

Calin, O. (2020). Neuromanifolds. *Deep Learning Architectures: A Mathematical Approach*, pages 465–504.

Constantinescu, V.-R. and Popescu, I. (2023). Interpolation property of shallow neural networks. *arXiv preprint arXiv:2304.10552*.

Draisma, J., Horobeţ, E., Ottaviani, G., Sturmfels, B., and Thomas, R. R. (2016). The euclidean distance degree of an algebraic variety. *Foundations of computational mathematics*, 16:99–149.

Finkel, B., Rodriguez, J. I., Wu, C., and Yahl, T. (2024). Activation thresholds and expressiveness of polynomial neural networks. *arXiv preprint arXiv:2408.04569*.

Fukushima, K. (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *IEICE Technical Report, A*, 62(10):658–665.

Hartshorne, R. (2013). *Algebraic geometry*, volume 52. Springer Science & Business Media.

Henry, N. W., Marchetti, G. L., and Kohn, K. (2024). Geometry of lightning self-attention: Identifiability and dimension. *arXiv preprint arXiv:2408.17221*.

Hou, L., Samaras, D., Kurc, T., Gao, Y., and Saltz, J. (2017). Convnets with smooth adaptive activation functions for regression. In *Artificial Intelligence and Statistics*, pages 430–439. PMLR.

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

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kohn, K. (2024). The geometry of the neuromanifold. *Collections*, 57(06).

Kohn, K., Merkh, T., Montúfar, G., and Trager, M. (2022). Geometry of linear convolutional networks. *SIAM Journal on Applied Algebra and Geometry*, 6(3):368–406.

Kohn, K., Montúfar, G., Shahverdi, V., and Trager, M. (2023). Function space and critical points of linear convolutional networks. *arXiv preprint arXiv:2304.05752*.

Kozhasov, K., Muniz, A., Qi, Y., and Sodomaco, L. (2023). On the minimal algebraic complexity of the rank-one approximation problem for general inner products. *arXiv preprint arXiv:2309.15105*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kubjas, K., Li, J., and Wiesmann, M. (2024). Geometry of polynomial neural networks. *arXiv preprint arXiv:2402.00949*.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Li, B., Tang, S., and Yu, H. (2019). Powernet: Efficient representations of polynomials and smooth functions by deep neural networks with rectified power units. *arXiv preprint arXiv:1909.05136*.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

López-Rubio, E., Ortega-Zamorano, F., Domínguez, E., and Muñoz-Pérez, J. (2019). Piecewise polynomial activation functions for feedforward neural networks. *Neural Processing Letters*, 50:121–147.

Milnor, J. W. (1963). *Morse theory*. Number 51. Princeton university press.

Shahverdi, V. (2024). Algebraic complexity and neurovariety of linear convolutional networks. *arXiv preprint arXiv:2401.16613*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Telgarsky, M. (2017). Neural networks and rational functions. In *International Conference on Machine Learning*, pages 3387–3393. PMLR.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

A ON THE DIFFERENTIAL OF THE PARAMETRIZATION

In this section, we prove technical results on the differential of the parametrization of polynomial CNNs. The regularity of the projectified parametrization follows immediately – see Theorem 4.5. We adhere to the convention from Section 4 and work with complex scalars. As discussed in Remark 4.3, the results hold, a posteriori, over $\mathbb{R}$ as well. Moreover, we denote by $\mathrm{J}_{\mathbf{w}}\varphi$ the differential of φ at $\mathbf{w}$. Since both the domain and co-domain of φ are vector spaces, the differential can be seen as a linear map $\mathrm{J}_{\mathbf{w}}\varphi\colon \mathbb{C}^{|\mathbf{k}|} \to \mathrm{Sym}_{r^{L-1}}(\mathbb{R}^{d_0})^{d_L}$. Moreover, for $L > 1$, we denote $\mathbf{w}' = (w_0, \dots, w_{L-2})$. Then $\varphi_{\mathbf{w}} = w_{L-1} \star_{s_{L-1}} \sigma_r(\varphi_{\mathbf{w}'})$, and by applying the Leibniz derivation rule we see that $\mathrm{J}_{\mathbf{w}}\varphi$ sends a tangent vector $\dot{\mathbf{w}} = (\dot{w}_0, \dots, \dot{w}_{L-1}) \in \mathbb{C}^{|\mathbf{k}|}$ to:

$$\dot{w}_{L-1} \star_{s_{L-1}} \sigma_r\left(\varphi_{\mathbf{w}'}\right) + r w_{L-1} \star_{s_{L-1}} \sigma_{r-1}\left(\varphi_{\mathbf{w}'}\right) \odot \left(\mathrm{J}_{\mathbf{w}'}\varphi\right)\left(\dot{\mathbf{w}}'\right), \tag{15}$$

where $\odot$ denotes the Hadamard product, i.e., the point-wise product between polynomial functions.

Lemma A.1. *Given $\lambda_0, \dots, \lambda_{L-1} \in \mathbb{C}$, the differential $\mathrm{J}_{\mathbf{w}}\varphi$ sends $(\lambda_0 w_0, \dots, \lambda_{L-1} w_{L-1})$ to:*

$$\left(\lambda_{L-1} + r\lambda_{L-2} + \cdots + r^{L-1}\lambda_0\right)\, \varphi_{\mathbf{w}}. \tag{16}$$

Proof. This follows immediately by substituting $\dot{w}_i = \lambda_i w_i$ for all i in Equation 15 and by reasoning inductively on the number of layers L. □

Proposition A.2. *Suppose that $w_i \neq 0$ for all i. Then the kernel of $\mathrm{J}_{\mathbf{w}}\varphi$ is generated, as a linear subspace of $\mathbb{C}^{|\mathbf{k}|}$, by $(0, \dots, 0, w_{L-2}, -r w_{L-1})$, $(0, \dots, 0, w_{L-3}, -r w_{L-2}, 0)$, …, $(w_0, -r w_1, 0, \dots, 0)$.*

Proof. We proceed by induction over the number of layers L. For $L = 1$, $\varphi_{\mathbf{w}}(x) = w_0 \star_{s_0} x$, which is linear in w_0 and does not vanish for all x since $w_0 \neq 0$. The differential is therefore injective, and its kernel is trivial.

Suppose now that $L > 1$. In order to compute the kernel $\mathrm{J}_{\mathbf{w}}\varphi$, we need to solve

$$\dot{w}_{L-1} \star_{s_{L-1}} \sigma_r\left(\varphi_{\mathbf{w}'}\right) + r w_{L-1} \star_{s_{L-1}} \sigma_{r-1}\left(\varphi_{\mathbf{w}'}\right) \odot \left(\mathrm{J}_{\mathbf{w}'}\varphi\right)\left(\dot{\mathbf{w}}'\right) = 0. \tag{17}$$

If $\dot{\mathbf{w}}'$ belongs to the kernel of $\mathrm{J}_{\mathbf{w}'}\varphi$, then since $\varphi_{\mathbf{w}'} \neq 0$ by hypothesis (see Corollary 4.2), we must have $\dot{w}_{L-1} = 0$. It follows from the inductive hypothesis that $\dot{w}$ is a linear combination of the tuples of filters in the statement. We therefore assume that $\left(\mathrm{J}_{\mathbf{w}'}\varphi\right)\left(\dot{\mathbf{w}}'\right) \neq 0$. Moreover, without loss of generality, we assume that $w_{L-1}[0] \neq 0$.

We will now decompose $\varphi_{\mathbf{w}}$ as a sum of CNNs with smaller filter size on the last layer. To this end, consider the tuple of filters:

$$\mathbf{w}^+ = (w_0, \dots, w_{L-2}, (w_{L-1}[0])), \qquad \mathbf{w}^- = (w_0, \dots, w_{L-2}, (0, w_{L-1}[1], \dots, w_{L-1}[k_{L-1} - 1])). \tag{18}$$

Then $\varphi_{\mathbf{w}} = \varphi_{\mathbf{w}^+} + \varphi_{\mathbf{w}^-}$. Now, let i be the minimal index such that $x[i]$ appears in $\varphi_{\mathbf{w}'}[0]$, seen as a scalar-valued polynomial in the input variable x. Recall the basic equivariance property of CNNs, meaning that shifting indices in the output variable is equivalent to shifting them in the input one (assuming the latter shift accounts for the strides). This implies that $x[i]$ does not appear in neither $\left(\mathrm{J}_{\mathbf{w}^-}\varphi\right)\left(\dot{\mathbf{w}}^-\right)$ nor $\varphi_{\mathbf{w}^-}$. Since $0 = \mathrm{J}_{\mathbf{w}}\varphi = \mathrm{J}_{\mathbf{w}^+}\varphi + \mathrm{J}_{\mathbf{w}^-}\varphi$, $x[i]$ does not appear in $\mathrm{J}_{\mathbf{w}^+}\varphi$ as well. Write:

$$\varphi_{\mathbf{w}'}[0] = x[i] f + g, \qquad \left(\mathrm{J}_{\mathbf{w}'}\varphi\right)\left(\dot{\mathbf{w}}'\right)[0] = x[i] a + b, \tag{19}$$

where f, g, a, b are polynomials such that $f \neq 0$ and $x[i]$ does not appear in g nor b. The (vanishing) terms containing $x[i]$ in $\mathrm{J}_{\mathbf{w}^+}\varphi$ have the form:

$$0 = \dot{w}_{L-1}[0]\left((x[i]f + g)^r - g^r\right) + r w_{L-1}[0]\left((x[i]f + g)^{r-1}(x[i]a + b) - g^{r-1}b\right). \tag{20}$$

By manipulating the above expression, we obtain:

$$(x[i]f + g)^{r-1}\left(\dot{w}_{L-1}[0](x[i]f + g) + r w_{L-1}[0](x[i]a + b)\right) = g^{r-1}(\dot{w}_{L-1}[0] g + r w_{L-1}[0] b). \tag{21}$$

The right-hand side of the above equation does not contain $x[i]$. Since $f \neq 0$, it follows that $\dot{w}_{L-1}[0](x[i]f + g) + r w_{L-1}[0](x[i]a + b) = 0$, implying:

$$\varphi_{\mathbf{w}'}[0] = x[i]a + b = \lambda\,(x[i]f + g) = \lambda\left(\mathrm{J}_{\mathbf{w}'}\varphi\right)\left(\dot{\mathbf{w}}'\right)[0], \qquad \lambda := -\frac{\dot{w}_{L-1}[0]}{w_{L-1}[0]}. \tag{22}$$

By looking at (the 0-th entry of) Equation 17, we deduce $\dot{w}_{L-1} = -r\lambda w_{L-1}$. By Lemma A.1, $\dot{\mathbf{w}}' = (\lambda_0 w_0, \dots, \lambda_{L-2} w_{L-2})$ for some $\lambda_i \in \mathbb{C} \setminus \{0\}$ such that $\lambda_{L-2} + r\lambda_{r-3} + \cdots + r^{L-2}\lambda_0 = \lambda$. But then

$$\dot{\mathbf{w}} = (\lambda_0 w_0, \dots, \lambda_{L-2} w_{L-2}, -r\lambda w_{L-1}) = \tag{23}$$

$$= \lambda(0, \dots, 0, w_{L-2}, -r w_{L-1}) + (\lambda_0 w_0, \dots, \lambda_{L-3} w_{L-3}, (\lambda_{L-2} - \lambda) w_{L-2}, 0). \tag{24}$$

Since $(\lambda_{L-2} - \lambda) + r\lambda_{L-3} + \cdots + r^{L-2}\lambda_0 = 0$, it follows from Lemma A.1 that the second summand of the above expression belongs to the kernel of $\mathrm{J}_{\mathbf{w}'}\varphi$, and the desired result follows from the inductive hypothesis. □

B ADDITIONAL PROOFS

B.1 Proof of Lemma 4.1

Proof. Since the domain of a convolution has higher dimension than the co-domain, we need to show that the rows $A_0, \dots, A_{d'-1}$ of the corresponding Toeplitz matrix are linearly independent. By induction on the filter size k, we can assume $w[0] \neq 0$. Suppose that $\sum_i a_i A_i = 0$ for some scalars $a_0, \dots, a_{d'-1}$. Since the first column of the Toeplitz matrix has only one non-vanishing entry, we deduce $a_0 = 0$. But then, since the s-th column is $(w[s-1], w[0], 0, \dots, 0)$, we similarly deduce that $a_1 = 0$, and so on. □

B.2 Proof of Proposition 4.3

Proof. Projective morphisms over algebraically-closed fields are closed (Hartshorne, 2013, II, §4, Theorem 4.9) and, in particular, have closed image. Therefore, $\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ is closed in $\mathbb{P}\mathrm{Sym}_{r^{L-1}}(\mathbb{C}^{d_0})^{d_L}$. Since φ is homogeneous, the neuromanifold coincides with the affine cone of its projectification. Since affine cones of varieties are varieties, $\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ is closed in $\mathrm{Sym}_{r^{L-1}}(\mathbb{C}^{d_0})^{d_L}$. □

B.3 Proof of Proposition 4.4

Proof. Newton's multinomial expansion yields:

$$(w \star_s x)[i]^r = \sum_{|\mathbf{a}|=r} \binom{r}{\mathbf{a}} \prod_{0 \le j < k} w[j]^{a_j} \prod_{0 \le j < k} x[is+j]^{a_j},$$

where $|\mathbf{a}| = a_0 + \cdots + a_{k-1}$ and $\binom{r}{\mathbf{a}} = \frac{r!}{a_0! \cdots a_{k-1}!}$. We construct $\widetilde{x}$ by indexing its coordinates via pairs $(i, \mathbf{a})$, where $1 \le i < d - k$ and $\mathbf{a}$ is a multi-index with $|\mathbf{a}| = r$, in some fixed order. The corresponding coordinate of $\tilde{x}$ is $\prod_j x[i+j]^{a_j}$. Similarly, $\tilde{w}$ is indexed by $\mathbf{a}$, with corresponding coordinate $\binom{r}{\mathbf{a}} \prod_j w[j]^{a_j}$, i.e., $\tilde{w}$ is a rescaling of the Veronese embedding of w of degree r. The claim follows immediately. □

B.4 Proof of Theorem 4.6

We first provide a general result on homogeneous polynomials.

Lemma B.1. *Assume that $r > 1$. Let $d, n \in \mathbb{N}$, $0 \le i < n$, and p be a multivariate homogeneous polynomial of degree d in n variables over a field of characteristic 0 (e.g., $\mathbb{R}$ or $\mathbb{C}$). Then the terms containing $x[i]$ in p^r determine p up to multiplicative scalar.*

Proof. Write $p = \sum_{0 \le j \le d} x[i]^j q_j$, where q_j is a homogeneous polynomial of degree $d - j$ not containing $x[i]$. Newton's multinomial expansion yields:

$$p^r = \sum_{|\mathbf{a}|=r} \binom{r}{\mathbf{a}} \prod_{0 \le j \le d} x[i]^{j a_j} q_j^{a_j} = \sum_{|\mathbf{a}|=r} \binom{r}{\mathbf{a}} x[i]^{\sum_j j a_j} \prod_{0 \le j \le d} q_j^{a_j}. \tag{25}$$

By following $t := \sum_{0 \le j \le d} j a_j$ in decreasing order, we can inductively recover q_j up to scalar in decreasing order w.r.t. j. More specifically, starting from $t = dr$, the coefficient of $x[i]^t$ is q_d^d, from which we recover q_d. Then, for $t = dr - 1$, the coefficient of $x[i]^t$ is $q_d^{r-1} q_{d-1}$, from which we can now recover q_{d-1}, and so on. □

We are now ready to prove Theorem 4.6.

Proof. We will prove the following equivalent statement on the non-projectified parametrization: if $\varphi_{\mathbf{w}}[0] = \varphi_{\mathbf{v}}[k]$ for some $0 < k < d_L$ and some $\mathbf{w}, \mathbf{v} \in \mathbb{C}^{|\mathbf{k}|}$ such that $w_i, v_i \neq 0$ for all $0 \leq i < L$, then $\mathbf{w}$ and $\mathbf{v}$ are related, up to rescaling, by a shift of their indices. To this end, we proceed by induction on L. For $L = 1$, the desired result follows from the fact that the linear map defined by a convolution determines the corresponding filter. Suppose now that $L > 1$ and that $\varphi_{\mathbf{w}}[0] = \varphi_{\mathbf{v}}[k]$. The inductive argument is similar in spirit to the one from the proof of Proposition A.2. Denote $\mathbf{w}' = (w_0, \dots, w_{L-2})$, and analogously for $\mathbf{v}'$. Then for all x:

$$\varphi_{\mathbf{w}}(x)[0] = \sum_{0 \leq j < k_{L-1}} w_{L-1}[j] \ (\varphi_{\mathbf{w}'}(x)[j])^r = \varphi_{\mathbf{v}}(x)[k] = \sum_{0 \leq j < k_{L-1}} v_{L-1}[j] \ (\varphi_{\mathbf{v}'}(x)[s_{L-1}k + j])^r. \tag{26}$$

Let i be the minimal index such that $x[i]$ appears in the above expression, seen as a scalar-valued polynomial in the input variable x. Moreover, let m and n be the minimal indices such that $w_{L-1}[m] \neq 0$ and $v_{L-1}[n] \neq 0$, respectively. Note that among all the summands in Equation 26, $x[i]$ appears only in $w_{L-1}[m] \ (\varphi_{\mathbf{w}'}(x)[m])^r$ and $v_{L-1}[n] \ (\varphi_{\mathbf{v}'}(x)[s_{L-1}k + n])^r$. Therefore, the terms involving $x[i]$ in these two summands must coincide. By Lemma B.1, $\varphi_{\mathbf{w}'}(x)[m] = \lambda \ \varphi_{\mathbf{v}'}(x)[s_{L-1}k + n]$ for some $\lambda \in \mathbb{C} \setminus \{0\}$. From the equivariance properties of convolutions, it follows that $\varphi_{\mathbf{w}'}(x)[j] = \lambda \ \varphi_{\mathbf{v}'}(x)[s_{L-1}k + n - m + j]$ for all $0 \leq j < d_{L-2}$. From the inductive hypothesis, for all $0 \leq t < L - 1$, w_t and v_t are related, up to rescaling, by a shift of their indices. Moreover, Equation 26 expands to:

$$\sum_{n \leq j < k_{L-1}} v_{L-1}[j] \ (\varphi_{\mathbf{v}'}(x)[s_{L-1}k + j])^r = \sum_{m \leq j < k_{L-1}} w_{L-1}[j]\lambda^r \ (\varphi_{\mathbf{v}'}(x)[s_{L-1}k + n - m + j])^r. \tag{27}$$

Again, in the above expression $x[i]$ appears only in the summands $v_{L-1}[n] \ (\varphi_{\mathbf{v}'}(x)[s_{L-1}k + n])^r$ and $w_{L-1}[m]\lambda^r \ (\varphi_{\mathbf{v}'}(x)[s_{L-1}k + n])^r$. Since these summands must coincide, we conclude that $v_{L-1}[n] = w_{L-1}[m]\lambda^r$. But then the indices of the sums in Equation 27 start from $n + 1$ and $m + 1$. By iterating this argument, we conclude that $v_{L-1}[n - m + j] = w_{L-1}[j]\lambda^r$ for all j, i.e., the filters of the last layer coincide up to rescaling and shifting the indices, as desired.

Lastly, note that the birationality statement on $\overline{\varphi}$ follows immediately. Indeed, the above argument shows that the fiber of $\overline{\varphi}$ at $\varphi_{\mathbf{w}}$ is not a singleton only if some filters in $\mathbf{w}$ are padded by zeros on the left or on the right, which is a negligible condition, i.e., it does not hold almost everywhere. The remaining fibers are characterized by shifting indices of filters, and are therefore finite. □

B.5 Proof of Corollary 4.7

Proof. Birational maps preserve the dimension. Therefore, the dimension of the projective neuromanifold is:

$$\dim(\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) = \dim(\mathbb{P}\mathbb{C}^{k_0} \times \cdots \times \mathbb{P}\mathbb{C}^{k_{L-1}}) = \sum_{0 \leq i < L} (k_i - 1) = |\mathbf{k}| - L. \tag{28}$$

Since $\dim(\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) = \dim(\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) - 1$, the dimension of the neuromanifold is $|\mathbf{k}| - L + 1$.

A birational linear (projective) map from an algebraic variety whose kernel does not intersect the variety preserves the degree. Since the projective neuromanifold is the image of a Segre–Veronese variety via a linear birational map, we have that $\deg(\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) = \deg(\mathcal{V}_{\mathbf{m},\mathbf{p}})$. The latter can be computed via Kozhasov et al. (2023, Proposition 6.11), as follows:

$$\deg(\mathcal{V}_{\mathbf{m},\mathbf{p}}) = (|\mathbf{k}| - L)! \prod_{0 \leq j < L} \frac{r^{(L-j-1)(k_j-1)}}{(k_j - 1)!}. \tag{29}$$

Finally, since the degree of a projective variety coincides with the one of its affine cone, we have $\deg(\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) = \deg(\mathbb{P}\mathcal{M}^{\mathbb{C}}_{\mathbf{d},\mathbf{k},\mathbf{s},r})$. □

B.6 Proof of Proposition 4.8

Proof. This follows from the regularity of the parametrization. Specifically, for an image of a real algebraic map, the relative boundary of the image inside its Zariski closure is contained in the branch locus of the complexification

of the map. Since the complexified projectified parametrization $\overline{\varphi}$ is regular (Theorem 4.5), it is unramified, i.e., the branching locus is empty. Therefore, $\mathbb{P}\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ is closed in the Zariski topology, and the same holds for its affine cone $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$. □

B.7 Proof of Proposition 4.10

Proof. From Section 4.1, we know that $\overline{\varphi}$ factors into the Segre–Veronese embedding $\nu_{\mathbf{m},\mathbf{p}}$, with $\mathbf{m} = (r^{L-1}, \dots, r, 1)$ and $\mathbf{p} = (k_0 - 1, \dots, k_{L-1} - 1)$, followed by a linear map. By Draisma et al. (2016, Corollary 6.1), linear projections of varieties of co-dimension ≥ 2 do not alter the generic Euclidean distance degree. Therefore, $\mathrm{gED}(\mathbb{P}\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) = \mathrm{gED}(\mathcal{V}_{\mathbf{m},\mathbf{p}})$. The generic Euclidean distance degree of the Segre–Veronese variety is given by Theorem 3.1, from which our formula follows via elementary algebraic manipulations. Note that, by definition, the Euclidean distance degree of a projective variety coincides with the one of its affine cone. Since φ is homogeneous, the affine cone of $\mathbb{P}\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$ is $\overline{\mathcal{M}}_{\mathbf{d},\mathbf{k},\mathbf{s},r}$, concluding the proof. □

Below, we provide a table with numerical values of the generic Euclidean distance degree.

r \ k	2	3	4	5	6
1	6	39	284	2205	17730
2	14	219	3772	68405	1277898
3	22	543	14684	417005	12186066
4	30	1011	37244	1439205	57202074
5	38	1623	75676	3699005	185917794
6	46	2379	134204	7933205	482134890

Table 1: The generic Euclidean distance degree of neuromanifolds with $L = 2$ and $k := k_1 = k_2$.

B.8 Proof of Proposition 4.11

Proof. Note that $\mathbf{w}$ is critical for $\mathrm{d}_A(\varphi_\bullet, u)^2$ if, and only if, $\varphi_{\mathbf{w}} - u$ is perpendicular, according to the scalar product induced by A, to the image of the differential of φ at $\mathbf{w}$. In other words, u must belong to the relative normal bundle of φ. By Proposition 4.7 and Theorem 4.5, such bundle consists, at a given $\varphi_{\mathbf{w}}$, of a finite number of affine subspaces of co-dimension $|\mathbf{k}| - L + 1 = \dim((\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}))$. If we constrain $\varphi_{\mathbf{w}}$ to be singular, then the restricted bundle has co-dimension $\dim(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r}) - \dim((\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s},r})_{\mathrm{sing}}) > 0$. Belonging to it is, therefore, a negligible condition in $u \in V$, which concludes the proof. □