# **Identifiability of Deep Polynomial Neural Networks**

## Konstantin Usevich, Ricardo Borsoi, Clara Dérand, Marianne Clausel

Université de Lorraine, CNRS, CRAN Nancy, F-54000, France firstname.lastname@univ-lorraine.fr

# **Abstract**

Polynomial Neural Networks (PNNs) possess a rich algebraic and geometric structure. However, their identifiability—a key property for ensuring interpretability—remains poorly understood. In this work, we present a comprehensive analysis of the identifiability of deep PNNs, including architectures with and without bias terms. Our results reveal an intricate interplay between activation degrees and layer widths in achieving identifiability. As special cases, we show that architectures with non-increasing layer widths are generically identifiable under mild conditions, while encoder-decoder networks are identifiable when the decoder widths do not grow too rapidly compared to the activation degrees. Our proofs are constructive and center on a connection between deep PNNs and low-rank tensor decompositions, and Kruskal-type uniqueness theorems. We also settle an open conjecture on the dimension of PNN's neurovarieties, and provide new bounds on the activation degrees required for it to reach the expected dimension.

## 1 Introduction

Neural network architectures which use polynomials as activation functions—polynomial neural networks (PNN)—have emerged as architectures that combine competitive experimental performance (capturing high-order interactions between input features) while allowing a fine grained theoretical analysis. On the one hand, PNNs have been employed in many problems in computer vision [1–3], image representation [4], physics [5] and finance [6], to name a few. On the other hand, the geometry of function spaces associated with PNNs, called neuromanifolds, can be analyzed using tools from algebraic geometry. Properties of such spaces, such as their dimension, shed light on the impact of a PNN architecture (layer widths and activation degrees) on the expressivity of feedforward, convolutional and self-attention PNN architectures [7–11]. They also determine the landscape of their loss function and the dynamics of their training process [7, 12, 13].

Moreover, PNNs are also closely linked to low-rank tensor decompositions [14–18], which play a fundamental role in the study of latent variable models due to their *identifiability* properties [19]. In fact, single-output 2-layer PNNs are equivalent to symmetric tensors [7]. Identifiability—whether the parameters and, consequently, the hidden representations of a NN can be determined from its response up to some equivalence class of trivial ambiguities such as permutations of its neurons—is a key question in NN theory [20–32]. Identifiability is critical to ensure interpretability in representation learning [33–35], to provably obtain disentangled representations [36], and in the study of causal models [37]. It is also critical to understand how the architecture affects the inference process and to support manipulation or "stitching" of pretrained models and representations [35, 38, 39]. Moreover, it has important links to learning and optimization of PNNs [40, 9, 13].

<sup>\*</sup>Corresponding author

<sup>&</sup>lt;sup>†</sup>M. Clausel is affiliated with Université de Lorraine, CNRS, IECL, Nancy.

The identifiability of deep PNNs is intimately linked to the dimension of their so-called *neurovarieties*: when it reaches the effective parameter count, the number of possible parametrizations is finite, which means the model is *finitely identifiable* and the neurovariety is said to be *non-defective*. In addition, many PNN architectures admit only a single parametrization (i.e., they are *globally identifiable*). This has been investigated for specific types of self-attention [9] and convolutional [8] layers, and feedforward PNNs without bias [11]. However, current results for feedforward networks only show that finite identifiability holds for very high activation degrees, or for networks with the same widths in every layer [11]. A standing conjecture is that this holds for any PNN with degrees at least quadratic and non-increasing layer widths [11], which parallels identifiability results of ReLU networks [29]. However, a general theory of identifiability of deep PNNs is still missing.

## 1.1 Our contribution

We provide a comprehensive analysis of the identifiability of deep PNNs considering monomial activation functions. We prove that an *L*-layer PNN is finitely identifiable if every 2-layer block composed by a pair of two successive layers is finitely identifiable for some subset of their inputs. This surprising result tightly links the identifiability of shallow and deep polynomial networks, which is a key challenge in the general theory of NNs. Moreover, our results reveal an intricate interplay between activation degrees and layer widths in achieving identifiability.

As special cases, we show that architectures with non-increasing layer widths (i.e., pyramidal nets) are generically identifiable, while encoder-decoder (bottleneck) networks are identifiable when the decoder widths do not grow too rapidly compared to the activation degrees. We also show that the minimal activation degrees required to render a PNN identifiable (which is equivalent to its *activation thresholds*) is only *linear* in the layer widths, compared to the quadratic bound in [11, Theorem 18]. These results not only settle but generalize conjectures stated in [11]. Moreover, we also address the case of PNNs with biases (which was overlooked in previous theoretical studies) by leveraging a *homogeneization* procedure.

Our proofs are constructive and are based on a connection between deep PNNs and partially symmetric canonical polyadic tensor decompositions (CPD). This allows us to leverage Kruskal-type uniqueness theorems for tensors to obtain identifiability results for 2-layer networks, which serve as the building block in the proof of the finite identifiability of deep nets, which is performed by induction. Our results also shed light on the geometry of the *neurovarieties*, as they lead to conditions under which its dimension reaches the expected (maximum) value.

## 1.2 Related works

**Polynomial NNs:** Several works studied PNNs from the lens of algebraic geometry using their associated *neuromanifolds* and *neurovarieties* [7] (in the emerging field of *neuroalgebraic geometry* [41]) and their close connection to tensor decompositions. Kileel et al. [7] studied the expressivity or feedforward PNNs in terms of the dimension of their neurovarieties. An analysis of the neuromanifolds for several architectures was presented in [10]. Conditions under which training losses do not exhibit bad local minima or spurious valleys were also investigated [13, 12, 42]. The links between training 2-layer PNNs and low-rank tensor approximation [13] as well as the biases gradient descent [43] have been established.

Recent work computed the dimensions of neuromanifolds associated with special types of self-attention [9] and convolutional [8] architectures, and also include identifiability results. For feedforward PNNs, finite identifiability was demonstrated for networks with the same widths in every layer [11], while stronger results are available for the 2-layer case with more general polynomial activations [44]. Finite identifiability also holds when the activation degrees are larger than a so-called *activation threshold* [11]. Recent work studied the singularities of PNNs with activations consisting of the sum of monomials with very high activation degrees [45]. PNNs are also linked to *factorization machines* [46]; this led to the development of efficient tensor-based learning algorithms [47, 48]. Note that other types of non-monomial polynomial-type activations [49, 50, 5, 51] have shown excellent performance; however, the geometry of these models is not well known.

NN identifiability: Many studies focused on the identifiability of 2-layer NNs with tanh, odd, and ReLU activation functions [20–23]. Moreover, algorithms to learn 2-layer NNs with unique parameter recovery guarantees have been proposed (see, e.g., [52, 53]), however, their extension to NNs with 3 or more layers is challenging and currently uses heuristics [54]. The identifiability of deep NNs

under weak genericity assumptions was first studied in the pioneering work of Fefferman [24] for the case of the tanh activation function through the study of its singularities. Recent work extended this result to more general sigmoidal activations [25, 26]. Various works focused on deep ReLU nets, which are piecewise linear [28]; they have been shown to be generically identifiable if the number of neurons per layer is non-increasing [29]. Recent work studied the local identifiability of ReLU nets [30–32]. Identifiability has also been studied for latent variable/causal modeling, leveraging different types of assumptions (e.g., sparsity, statistical independence, etc.) [55–60]. Note that although some of these works tackle deep NNs, their proof techniques are completely different from our approach and do not apply to the case of polynomial activation functions.

**Tensors and NNs:** Low-rank tensor decompositions had widespread practical impact in the compression of NN weights [61–65]. Moreover, their properties also played a key role in the theory of NNs [18]. This includes the study of the expressivity of convolutional [66] and recurrent [67, 68] NNs, and the sample complexity of reinforcement learning parametrized by low-rank transition and reward tensors [69, 70]. The decomposability of low-rank symmetric tensors was also paramount in establishing conditions under which 2-layer NNs can (or cannot [71]) be learned in polynomial time and in the development of algorithms with identifiability guarantees [52, 72, 73]. It was also used to study identifiability of some deep *linear* networks [74]. However, the use of tensor decompositions in the studying the identifiability of deep *nonlinear* networks has not yet been investigated.

# 2 Setup and background

# 2.1 Polynomial neural networks: with and without bias

Polynomial neural networks are functions  $\mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$  represented as feedforward networks with bias terms and activation functions of the form  $\rho_r(\cdot) = (\cdot)^r$ . Our results hold for both the real and complex valued case ( $\mathbb{F} = \mathbb{R}, \mathbb{C}$ ), thus, and we prefer to keep the real notation for simplicity. Note that we allow the activation functions to have a different degree  $r_\ell$  for each layer.

**Definition 1** (PNN). A polynomial neural network (PNN) with biases and architecture  $(d = (d_0, d_1, \ldots, d_L), r = (r_1, \ldots, r_{L-1}))$  is a map  $\mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$  given by a feedforward neural network

$$PNN_{\boldsymbol{d},\boldsymbol{r}}[\boldsymbol{\theta}] = PNN_{\boldsymbol{r}}[\boldsymbol{\theta}] := f_L \circ \rho_{r_{L-1}} \circ f_{L-1} \circ \rho_{r_{L-2}} \circ \cdots \circ \rho_{r_1} \circ f_1,$$
 (1)

where  $f_i(\mathbf{x}) = \mathbf{W}_i \mathbf{x} + \mathbf{b}_i$  are affine maps, with  $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$  being the weight matrices and  $\mathbf{b}_i \in \mathbb{R}^{d_i}$  the biases, and the activation functions  $\rho_r : \mathbb{R}^d \to \mathbb{R}^d$ , defined as  $\rho_r(\mathbf{z}) := (z_1^r, \dots, z_d^r)$  are monomial. The parameters  $\boldsymbol{\theta}$  are given by the entries of the weights  $\mathbf{W}_i$  and biases  $\mathbf{b}_i$ , i.e.,

$$\theta = (w, b), \ w = (W_1, W_2, \dots, W_L), \ b = (b_1, b_2, \dots, b_L).$$
 (2)

The vector of degrees r is called the activation degree of  $PNN_r[\theta]$  (we often omit the subscript d if it is clear from the context).

PNNs are algebraic maps and are polynomial vectors, where the total degree is  $r_{total} = r_1 \cdots r_{L-1}$ , that is, they belong to the polynomial space  $(\mathscr{P}_{d,r_{total}})^{\times d_L}$ , where  $\mathscr{P}_{d,r}$  denotes the space of d-variate polynomials of degree  $\leq r$ . Most previous works analyzed the simpler case of PNNs without bias, which we refer to as *homogeneous*. Due to its importance, we consider it explicitly.

**Definition 2** (hPNN). A PNN is said to be a **homogenous** PNN (hPNN) when it has no biases ( $\mathbf{b}_{\ell} = 0$  for all  $\ell = 1, ..., L$ ), and is denoted as

$$hPNN_{\boldsymbol{d},\boldsymbol{r}}[\boldsymbol{w}] = hPNN_{\boldsymbol{r}}[\boldsymbol{w}] := \boldsymbol{W}_{L} \circ \rho_{r_{L-1}} \circ \boldsymbol{W}_{L-1} \circ \rho_{r_{L-2}} \circ \cdots \circ \rho_{r_{1}} \circ \boldsymbol{W}_{1}. \tag{3}$$

Its parameter set is given by  $\mathbf{w} = (\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_L)$ .

It is well known that such PNNs are in fact homogeneous polynomial vectors and belong to the polynomial space  $(\mathscr{H}_{d_0,r_{total}})^{\times d_L}$ , where  $\mathscr{H}_{d,r} \subset \mathscr{P}_{d,r}$  denotes the space of homogeneous d-variate polynomials of degree r. hPNNs are also naturally linked to tensors and tensor decompositions, whose properties can be used in their theoretical analysis.

**Example 3** (Running example). Consider an hPNN with L = 2, r = (2) and d = (3, 2, 2). In such a case the parameter matrices are given as

$$m{W}_2 = egin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}, \quad m{W}_1 = egin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix},$$

and the hPNN  $p = \text{hPNN}_r[w]$  is a vector polynomial that has expression

$$\boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{W}_2 \rho_2(\boldsymbol{W}_1 \boldsymbol{x}) = \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} (a_{11}x_1 + a_{12}x_2 + a_{13}x_3)^2 + \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix} (a_{21}x_1 + a_{22}x_2 + a_{23}x_3)^2.$$

the only monomials that can appear are of the form  $x_1^i x_2^j x_3^k$  with i + j + k = 2 thus p is a vector of degree-2 homogeneous polynomials in 3 variables (in our notation,  $p \in (\mathcal{H}_{3,2})^2$ ).

## 2.2 Equivalent PNN representations

It is known that the PNNs admit equivalent representations (i.e., several parameters  $\theta$  lead to the same function). Indeed, for each hidden layer we can (a) permute the hidden neurons, and (b) rescale the input and output to each activation function since for any  $a \neq 0$ ,  $(at)^r = a^r t^r$ . This transformation leads to a different set of parameters that leave the PNN unchanged. We can characterize all such equivalent representations in the following lemma (provided in [7] for the case without biases).

**Lemma 4.** Let  $\text{PNN}_{d,r}[\theta]$  be a PNN with  $\theta$  as in (2). Let also  $D_{\ell} \in \mathbb{F}^{d_{\ell} \times d_{\ell}}$  be any invertible diagonal matrices and  $P_{\ell} \in \mathbb{Z}^{d_{\ell} \times d_{\ell}}$  ( $\ell = 1, \dots, L-1$ ) be permutation matrices, and define the transformed parameters as

$$oldsymbol{W}_\ell' \leftarrow oldsymbol{P}_\ell oldsymbol{D}_\ell oldsymbol{W}_\ell oldsymbol{D}_{\ell-1}^{-r_{\ell-1}} oldsymbol{P}_{\ell-1}^\mathsf{T} \,, \qquad oldsymbol{b}_\ell' \leftarrow oldsymbol{P}_\ell oldsymbol{D}_\ell oldsymbol{b}_\ell \,,$$

with  $P_0 = D_0 = I$  and  $P_L = D_L = I$ . Then the modified parameters  $W'_{\ell}, b'_{\ell}$  define exactly the same network, i.e.  $\text{PNN}_{d,r}[\theta] = \text{PNN}_{d,r}[\theta']$  for the parameter vector

$$\theta' = ((W_1', W_2', \dots, W_L'), (b_1', b_2', \dots, b_L')).$$

If  $\theta$  and  $\theta'$  are linked with such a transformation, they are called equivalent (denoted  $\theta \sim \theta'$ ).

**Example 5** (Example 3, continued). *In Example 3 we can take any*  $\alpha$ ,  $\beta \neq 0$  *to get* 

$$\text{hPNN}_{\boldsymbol{d},\boldsymbol{r}}[\boldsymbol{w}](\boldsymbol{x}) = \begin{bmatrix} \alpha^{-2}b_{11} \\ \alpha^{-2}b_{21} \end{bmatrix} (\alpha a_{11}x_1 + \alpha a_{12}x_2 + \alpha a_{13}x_3)^2 + \begin{bmatrix} \beta^{-2}b_{12} \\ \beta^{-2}b_{22} \end{bmatrix} (\alpha a_{21}x_1 + \alpha a_{22}x_2 + \alpha a_{23}x_3)^2.$$

which correspond to rescaling rows of  $W_1$  and corresponding columns of  $W_2$ . If we additionally permute them, we get  $W_1' = PDW_1, W_2' = W_2D^{-2}P^T$  with  $D = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$  and  $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ .

This characterization of equivalent representations allows us to define when a PNN is unique.

**Definition 6** (Unique and finite-to-one representation). The PNN  $p = \text{PNN}_{d,r}[\theta]$  (resp. hPNN  $p = \text{hPNN}_{d,r}[w]$ ) with parameters  $\theta$  (resp. w) is said have a **unique** representation if every other representation satisfying  $p = \text{PNN}_{d,r}[\theta']$  (resp.  $p = \text{hPNN}_{d,r}[w']$ ) is given by an equivalent set of parameters, i.e.,  $\theta' \sim \theta$  (resp.  $w' \sim w$ ) in the sense Lemma 4 (i.e., they can be obtained from the permutations and elementwise scalings in Lemma 4).

Similarly, a PNN  $p = \text{PNN}_{d,r}[\theta]$  (resp.  $hPNN p = \text{hPNN}_{d,r}[w]$ ) is called **finite-to-one** if it admits only finitely many non-equivalent representations, that is, the set  $\{\theta' : \text{PNN}_{d,r}[\theta'] = p\}$  (resp.  $\{w' : \text{hPNN}_{d,r}[w'] = p\}$ ) contains finitely many non-equivalent parameters.

**Example 7** (Example 5, continued). Thanks to links with tensor decompositions, it is known that the hPNN in Example 3 is unique if  $\mathbf{W}_2$  is invertible and  $\mathbf{W}_1$  full row rank (rank 2).

# 2.3 Identifiability and link to neurovarieties

An immediate question is which PNN/hPNN architectures are expected to admit only a single (or finitely many) non-equivalent representations? This question can be formalized using the notions of **global** and **finite identifiability**, which considers a general set of parameters.

**Definition 8** (Global and finite identifiability). The PNN (resp. hPNN) with architecture (d, r) is said to be **globally identifiable** if for a general choice of  $\theta = (w, b) \in \mathbb{R}^{\sum d_{\ell}(d_{\ell-1}+1)}$ , (resp.  $w \in \mathbb{R}^{\sum d_{\ell}d_{\ell-1}}$ ) (i.e., for all choices of parameters except for a set of Lebesgue measure zero), the network  $\text{PNN}_{d,r}[\theta]$  (resp.  $\text{hPNN}_{d,r}[w]$ ) has a unique representation.

Similarly, the PNN (resp. hPNN) with architecture (d, r) is said to be **finitely identifiable** if for a general choice of  $\theta$ , (resp. w) the network  $\text{PNN}_{d,r}[\theta]$  (resp.  $\text{hPNN}_{d,r}[w]$ ) is finite-to-one (i.e., it admits only finitely many non-equivalent representations).

In the following, we use the term "identifiable" to refer to finite identifiability unless stated otherwise. Note also that the notion of finite identifiability is much stronger than the related notion of local identifiability (i.e., a model being identifiable only in a neighborhood of a parameterization).

**Example 9** (Example 7, continued). From Example 7, we see that the hPNN architecture with d = (3, 2, 2), r = (2) is identifiable due to the fact that generic matrices  $W_1$  and  $W_2$  are full rank.

Note that Definition 8 excludes a set of parameters of Lebesgue measure zero. Thus, for an identifiable architecture such as the one mentioned in Example 9, there exists rare sets of pathological parameters for which the hPNN is non-unique (e.g., weight matrices containing collinear rows).

With some abuse of notation, let  $\operatorname{hPNN}_{d,r}[\cdot]$  be the map taking w to  $\operatorname{hPNN}_{d,r}[w]$ . Then the image of  $\operatorname{hPNN}_{d,r}[\cdot]$  is called a *neuromanifold*, and the *neurovariety*  $\mathcal{V}_{d,r}$  is defined as its closure in the Zariski topology<sup>3</sup>. The study of neurovarieties and their properties is a topic of recent interest [7, 41, 11, 10]. More details are given in Appendix D. An important property for our case is the link between identifiability of an hPNN and the dimension of its neurovariety.

**Proposition 10.** The architecture  $\text{hPNN}_{d,r}[\cdot]$  is finitely identifiable if and only if the neurovariety has the expected (maximal possible) dimension  $\dim \mathcal{V}_{d,r} = \sum_{\ell=1}^L d_\ell d_{\ell-1} - \sum_{\ell=1}^{L-1} d_\ell$ . In such case,  $\mathcal{V}_{d,r}$  is said to be **nondefective**.

Our results concern finite identifiablity of PNN and hPNN architectures, which provides essential theoretical support for the interpretability of their representations.

# 3 Main results

## 3.1 Main results on the identifiability of deep hPNNs

Although several works have studied the identifiability of 2-layer NNs, tackling the case of deep networks is significantly harder. However, when we consider the opposite statement, i.e., the *non-identifiability* of a network, it is much easier to show such connection: in a deep network with L>2 layers, the lack of identifiability of any 2-layer subnetwork (formed by two consecutive layers) clearly implies that the full network is not identifiable. What our main result shows is that, surprisingly, under mild additional conditions the converse is also true for hPNNs: if the every 2-layer subnetwork is identifiable for some subset of their inputs, then the full network is identifiable as well. This is formalized in the following theorem.

**Theorem 11** (Localization theorem). Let  $((d_0,\ldots,d_L),(r_1,\ldots,r_{L-1}))$  be the hPNN format. For  $\ell=0,\ldots,L-2$  denote  $\widetilde{d}_\ell:=\min\{d_0,\ldots,d_\ell\}$ . Then the following holds true: if for all  $\ell=1,\ldots,L-1$  the two-layer architecture  $\operatorname{hPNN}_{(\widetilde{d}_{\ell-1},d_\ell,d_{\ell+1}),r_\ell}[\cdot]$  is finitely identifiable, then the L-layer architecture  $\operatorname{hPNN}_{d,r}[\cdot]$  is finitely identifiable as well.

The technical proofs are relegated to the appendices. This key result shows a strict equivalence between the finite identifiability of shallow and deep hPNNs. However, as we move into the deeper layers, the identifiability conditions required by Theorem 11 are slightly stricter than in the shallow case, since the number of inputs is reduced to  $\widetilde{d}_{\ell}$ . This can lead to a requirement of larger activation degrees to guarantee identifiability compared to the shallow case.

Theorem 11 allows us to derive identifiability conditions for hPNNs using the link between 2-layer hPNNs and partially symmetric tensor decompositions and their generic uniqueness based on classical Kruskal-type conditions. We use the following sufficient condition for the identifiability of shallow networks.

**Proposition 12.** Let  $m, d \geq 2$ ,  $n \geq 1$  be the layer widths and  $r \geq 2$  such that

$$r \ge \frac{2d - \min(n, d)}{\min(d, m) - 1}.\tag{4}$$

Then the 2-layer hPNN with architecture ((m, d, n), r) is globally identifiable.

**Remark 13.** If the above result holds for  $\ell = 1, ..., L-1$  with  $m = \widetilde{d}_{\ell-1}$ ,  $d = d_{\ell}$ ,  $n = d_{\ell+1}$  and  $r = r_{\ell}$ , then Theorem 11 implies that the L-layer hPNN is identifiable for the architecture (d, r).

<sup>&</sup>lt;sup>3</sup>i.e., the smallest algebraic variety that contains the image of the map  $hPNN_{d,r}[\cdot]$ .

**Remark 14.** Note that for the single output case  $d_L = 1$ , Equation (4) means the activation degree in the last layer must satisfy  $r_{L-1} \geq 3$ , in contrast to  $r_{\ell} \geq 2$  for  $\ell < L - 1$ .

**Remark 15** (Our bounds are constructive). We note that the condition (4) for identifiability is not the best possible (and can be further improved using much stronger results on generic uniqueness of decompositions, see e.g., [75, Corollary 37]). However, the bound (4) is constructive, and we can use standard polynomial-time tensor algorithms to recover the parameters of the 2-layer hPNN.

# 3.2 Implications for specific architectures

Proposition 12 has direct implications for the finite identifiability of several architectures of practical interest, including pyramidal and bottleneck networks, and for the activation thresholds of hPNNs, as shown in the following corollaries.

**Corollary 16** (Pyramidal hPNNs are always identifiable). The hPNNs with architectures containing non-increasing layer widths  $d_0 \ge d_1 \ge \cdots d_{L-1} \ge 2$ , except possibly for  $d_L \ge 1$  are finitely identifiable for any degrees satisfying

(i) 
$$r_1, \ldots, r_{L-1} \ge 2$$
 if  $d_L \ge 2$ ; or (ii)  $r_1, \ldots, r_{L-2} \ge 2$ ,  $r_{L-1} \ge 3$  if  $d_L \ge 1$ .

Note that, due to the connection between the identifiability of hPNNs and the neurovarieties presented in Proposition 10, a direct consequence of Corollary 16 is that the neurovariety  $\mathcal{V}_{d,r}$  has expected dimension. This settles a recent conjecture presented in [11, Section 4]. This implication is explained in detail in Appendix D.

Instead of seeking conditions on the layer widths for a fixed (or minimal) degree, a complementary perspective is to determine what are the smallest degrees  $r_{\ell}$  such that a given architecture d is finitely identifiable. Following the terminology introduced in [11], we refer to those values as the *activation thresholds for identifiability* of an hPNN. An upper bound is given in the following corollary:

**Corollary 17** (Activation thresholds for identifiability). For fixed layer widths  $\mathbf{d} = (d_0, \dots, d_L)$  with  $d_\ell \geq 2$ ,  $\ell = 0, \dots, L-1$ , the hPNNs with architectures  $(\mathbf{d}, (r_1, \dots, r_{L-1}))$  are finitely identifiable for any degrees satisfying

$$r_{\ell} \geq 2d_{\ell} - 1$$
.

Note that due to Proposition 10, the result in this corollary implies that the neurovariety  $\mathcal{V}_{d,r}$  has expected dimension. This means that  $(2d_{\ell}-1)$  is also a universal upper bound to the so-called *activation thresholds* for hPNN expressiveness introduced in [11]. The existence of such activation thresholds was conjectured in [7] and recently proved in [11, Theorem 18], but the for a *quadratic* in  $d_{\ell}$  bound (our bound is *linear*).

**Remark 18** (Admissible layer sizes). The possible layer sizes in a deep network are tightly linked with the degree of the activation. For example, for  $r_{\ell}=2$ , identifiability is impossible if  $d_{\ell}>\frac{d_{\ell-1}(d_{\ell-1}+1)}{2}$  (for general  $r_{\ell}$ , a similar bound  $O(d_{\ell-1}^{r_{\ell}})$  follows from a link with tensor decompositions [76]). Therefore, to allow for larger layer widths, we need to have higher-degree activations.

It is enlightening to consider the admissible layer widths when taking into account the joint effect of layer widths and degrees. By doing this, Proposition 12 can be leveraged to yield identifiability conditions for the case of bottleneck networks, as illustrated in the following corollary.

Corollary 19 (Identifiability of bottleneck hPNNs). Consider the "bottleneck" architecture with

$$d_0 \ge d_1 \ge \dots \ge d_b \le d_{b+1} \le \dots \le d_L$$

and  $d_b \geq 2$ . Suppose that  $r_1, \ldots, r_b \geq 2$  and that the decoder part satisfies  $\frac{d_\ell}{r_\ell} \leq d_b - 1$  for  $\ell \in \{b+1, \ldots, L-1\}$ . Then the bottleneck hPNN is finitely identifiable.

This shows that encoder-decoder hPNNs architectures are identifiable under mild conditions on the layer widths and decoder degrees, providing a polynomial networks-based counterpart to previous studies that analyzed linear autoencoders [77, 78].

Note that the width of the bottleneck layer  $d_b$  constrains the entire decoder part of the architecture: the degrees  $r_\ell, \ell \geq b$  are constrained according to the width  $d_b$ . The presence of bottlenecks has also been shown to affect the expressivity of hPNNs in [7, Theorem 19]: for  $d_b = 2d_0 - 2$  there exists a number of layers L such that for  $r_\ell \geq 2$  and  $d_0 \geq 2$ , the hPNN neurovariety is non-filling (i.e., its dimension never reaches that of the ambient space) for any choice of widths  $d_1, \ldots, d_{b-1}, d_{b+1}, \ldots, d_L$ .

## 3.3 PNNs with biases

The identifiability of general PNNs (with biases) can be studied via the properties of hPNNs. The simplest idea is *truncation* (i.e., taking only higher-order terms of the polynomials), which eliminates biases from PNNs. Such an approach was already taken in [44] for shallow PNNs with general polynomial activation, and is described in Appendix E. We will follow a different approach based on the well-known idea of **homogenization**: we transform a PNN to an equivalent hPNN with structured parameters keeping the information about biases at the expense of increasing the layer widths. Our key result is to show how this can be used to study the identifiability of PNNs with bias terms. The following correspondence is well-known.

**Definition 20** (Homogenization). There is a one-to-one mapping between polynomials in d variables of degree r and homogeneous polynomials of the same degree in d+1 variables. We denote this mapping  $\mathscr{P}_{d,r} \to \mathscr{H}_{d+1,r}$  by  $homog(\cdot)$ , and it acts as follows: for every polynomial  $p \in \mathscr{P}_{d,r}$ ,  $\widetilde{p} = homog(p) \in \mathscr{H}_{d+1,r}$  (that is  $\widetilde{p}(x_1, \ldots, x_d, x_{d+1})$ ) is the unique homogeneous polynomial in d+1 variables such that

$$\widetilde{p}(x_1,\ldots,x_d,1)=p(x_1,\ldots,x_d).$$

**Example 21.** For the polynomial  $p \in \mathscr{P}_{2,2}$  in variables  $(x_1, x_2)$  given by

$$p(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2 + ex_1 + fx_2 + g,$$

its homogenization  $\widetilde{p} = \text{homog}(p) \in \mathcal{H}_{3,2}$  in 3 variables  $(x_1, x_2, x_3)$  is

$$\widetilde{p}(x_1, x_2, x_3) = ax_1^2 + bx_1x_2 + cx_2^2 + ex_1x_3 + fx_2x_3 + gx_3^3$$

and we can verify that  $\widetilde{p}(1, x_1, x_2) = p(x_1, x_2)$ .

Similarly, we extend homogenization to polynomial vectors, which gives the following.

**Example 22.** Let  $f(x) = W_2 \rho_{r_1}(W_1 x + b_1) + b_2$ , and define extended matrices as

$$\widetilde{\boldsymbol{W}}_1 = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{b}_1 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{(d_1+1) \times (d_0+1)}, \quad \widetilde{\boldsymbol{W}}_2 = \begin{bmatrix} \boldsymbol{W}_2 & \boldsymbol{b}_2 \end{bmatrix} \in \mathbb{R}^{d_2 \times (d_1+1)}$$

Then its homogenization  $\widetilde{f} = \text{homog}(f)$  is an hPNN of format  $(d_0 + 1, d_1 + 1, d_2)$ 

$$\widetilde{f}(\widetilde{\boldsymbol{x}}) = \widetilde{\boldsymbol{W}}_2 \rho_{r_1} \left( \widetilde{\boldsymbol{W}}_1 \widetilde{\boldsymbol{x}} \right)$$

where 
$$\widetilde{x} = [x_0, x_1, \dots, x_{d_0}, x_{d_0+1}]^\mathsf{T}$$
, so that  $\widetilde{f}(x_1, \dots, x_{d_0}, 1) = f(x_1, \dots, x_{d_0})$ .

The construction in Example 22 similar to the well-known idea of augmenting the network with an artificial (constant) input. The following proposition generalizes this example to the case of multiple layers, by "propagating" the constant input.

**Proposition 23.** Fix the architecture  $\mathbf{r} = (r_1, \dots, r_L)$  and  $\mathbf{d} = (d_0, \dots, d_L)$ . Then a polynomial vector  $\mathbf{p} \in (\mathscr{P}_{d_0, r_{total}})^{\times d_L}$  admits a PNN representation  $\mathbf{p} = \text{PNN}_{\mathbf{d}, \mathbf{r}}[(\mathbf{w}, \mathbf{b})]$  with  $(\mathbf{w}, \mathbf{b})$  as in (2) if and only if its homogenization  $\tilde{\mathbf{p}} = \text{homog}(\mathbf{p})$  admits an hPNN decomposition for the same activation degrees  $\mathbf{r}$  and extended  $\tilde{\mathbf{d}} = (d_0 + 1, \dots, d_{L-1} + 1, d_L)$ ,  $\tilde{\mathbf{p}} = \text{hPNN}_{\tilde{\mathbf{d}}, \mathbf{r}}[\tilde{\mathbf{w}}]$ ,  $\tilde{\mathbf{w}} = (\tilde{\mathbf{W}}_1, \dots, \tilde{\mathbf{W}}_L)$ , with matrices given as

$$\widetilde{\boldsymbol{W}}_{\ell} = \begin{cases} \begin{bmatrix} \boldsymbol{W}_{\ell} & \boldsymbol{b}_{\ell} \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{(d_{\ell}+1) \times (d_{\ell-1}+1)}, & \ell < L, \\ \begin{bmatrix} \boldsymbol{W}_{L} & \boldsymbol{b}_{L} \end{bmatrix} \in \mathbb{R}^{(d_{L}) \times (d_{L-1}+1)}, & \ell = L. \end{cases}$$

That is, PNNs are in one-to-one correspondence to hPNNs with increased number of inputs and structured weight matrices.

**Uniqueness of PNNs from homogenization:** An important consequence of homogenization is that the uniqueness of the homogeneized hPNN implies the uniqueness of the original PNN with bias terms, which is a key result to support the application of our identifiability results to general PNNs.

**Proposition 24.** If  $\text{hPNN}_r[\widetilde{\boldsymbol{w}}]$  from Proposition 23 is unique (resp. finite-to-one) as an hPNN (without taking into account the structure), then the original PNN representation  $\text{PNN}_r[(\boldsymbol{w}, \boldsymbol{b})]$  is unique (resp. finite-to-one).

The proposition follows from the fact that we can always fix the permutation ambiguity for the "artificial" input.

**Remark 25.** Despite the one-to-one correspondence, we cannot simply apply identifiability results from the homogeneous case, because the matrices  $\widetilde{W}_{\ell}$  are structured (they form a set of measure zero inside  $\mathbb{R}^{(d_{\ell}+1)\times(d_{\ell-1}+1)}$ ).

However, we can prove that the identifiability of the hPNN implies the identifiability of the PNN.

**Lemma 26.** Let the 2-layer hPNN architecture be finitely (resp. globally) identifiable for  $((d_0 + 1, d_1 + 1, d_2), r_1)$ . Then the PNN architecture with widths  $(d_0, d_1, d_2)$  and degree  $r_1$  is also finitely (resp. globally) identifiable.

Using Lemma 26 and specializing the proof of Theorem 11, we obtain the following result:

**Proposition 27.** Let  $((d_0, \ldots, d_L), (r_1, \ldots, r_{L-1}))$  be the PNN format. For  $\ell = 0, \ldots, L-2$  denote  $\widetilde{d}_{\ell} = \min\{d_0, \ldots, d_{\ell}\}$ . Then the following holds true: If for all  $\ell = 1, \ldots, L-1$  the two layer architecture  $\operatorname{hPNN}_{(\widetilde{d}_{\ell-1}+1, d_{\ell+1}, d_{\ell+1}), r_{\ell}}[\cdot]$  is finitely identifiable, then the L-layer PNN with architecture (d, r) is finitely identifiable as well.

In particular, we have the following bounds for generic uniqueness.

**Corollary 28.** Let  $((d_0,\ldots,d_L),(r_1,\ldots,r_{L-1}))$  be such that  $d_\ell \geq 1$ , and  $r_\ell \geq 2$  satisfy

$$r_{\ell} \ge \frac{2(d_{\ell}+1) - \min(d_{\ell}+1, d_{\ell+1})}{\min(d_{\ell}, \tilde{d}_{\ell-1})},$$

then the L-layer PNN with architecture (d, r) is finitely identifiable (and globally identifiable if L = 2).

**Remark 29.** For the case of general PNNs with bias, similar conclusions to the hPNN case hold. For fixed layer widths  $d_{\ell} \geq 1$ , the activation threshold for a PNN architecture  $(\boldsymbol{d}, \boldsymbol{r})$  becomes  $r_{\ell} \geq 2d_{\ell} + 1$ . Also, pyramidal PNNs are identifiable in degree 2.

A remarkable feature of PNNs with bias is that they can be identifiable even for architectures with layers containing a single hidden neuron: for  $d_{\ell}=1$  and  $d_{\ell+1}\geq 2$  and/or  $\widetilde{d}_{\ell-1}=1$ , the condition in Corollary 28 is still satisfied when  $r_{\ell}\geq 2$ .

# 4 Proofs and main tools

Our main results in Theorem 11 translates the identifiability conditions of deep hPNNs into those of shallow hPNNs. Our results are strongly related to the decomposition of partially symmetric tensors (we review basic facts about tensors and tensors decompositions and recall their connection between to hPNNs in later subsections). More details are provided in the appendices, and we list key components of the proof below.

## 4.1 Identifiability of deep PNNs: necessary conditions

**Increasing hidden layers breaks uniqueness.** The key insight is that if we add to any architecture a neuron in any hidden layer, then the uniqueness of the hPNN is not possible, which is formalized as following lemma (whose proof is based, in its turn, on tensor decompositions).

**Lemma 30.** Let  $p = \text{hPNN}_r[w]$  be an hPNN of format  $(d_0, \dots, d_\ell, \dots, d_L)$ . Then for any  $\ell$  there exists an infinite number of representations of hPNNs  $p = \text{hPNN}_r[w]$  with architecture  $(d_0, \dots, d_\ell + 1, \dots, d_L)$ . In particular, the augmented hPNN is not unique (or finite-to-one).

**Internal features of a unique hPNN are linearly independent.** This is an easy consequence of Lemma 30 (as linear dependence would allow for pruning neurons).

**Lemma 31.** For  $d = (d_0, \dots, d_L)$ , let  $p = \text{hPNN}_r[w]$  have a unique (or finite-to-one) L-layers decomposition. Consider the output at any  $\ell$ -th internal level  $\ell < L$  after the activations

$$\boldsymbol{q}_{\ell}(\boldsymbol{x}) = \rho_{r_{\ell}} \circ \boldsymbol{W}_{\ell} \circ \cdots \circ \rho_{r_{1}} \circ \boldsymbol{W}_{1}(\boldsymbol{x}). \tag{5}$$

Then the elements of  $q_{\ell}(x) = [q_{\ell,1}(x) \cdots q_{\ell,d_{\ell}}(x)]^{\mathsf{T}}$  are linearly independent polynomials.

**Identifiability for hPNNs and Kruskal rank.** Identifiability of 2-layer hPNNs, or equivalently uniqueness of CPD is strongly related to the concept of Kruskal rank of a matrix that we define below.

**Definition 32.** The Kruskal rank of a matrix A (denoted krank $\{A\}$ ) is the maximal number k such that any k columns of A are linearly independent.

This is in contrast with the usual rank which requires that there exists k linearly independent columns. Therefore  $\operatorname{krank}\{A\} \leq \operatorname{rank}\{A\}$ . Note that  $\operatorname{krank}\{A\} \leq 1$  means that the matrix A has at least two columns that are linearly dependent (proportional). Using the notion of Kruskal rank, we can state a necessary condition on weight matrices for identifiability of hPNNs, which is a generalization of the well-known necessary condition for the uniqueness of CPD tensor decompositions (6) (i.e., shallow networks), and is a corollary of Lemma 30 and Lemma 31.

**Proposition 33.** As in Lemma 31, let the widths be  $\mathbf{d} = (d_0, \dots, d_L)$ , and  $\mathbf{p} = \text{hPNN}_{\mathbf{r}}[\mathbf{w}]$  have a unique (or finite-to-one) L-layers decomposition. Then we have that for all  $\ell = 1, \dots, L-1$ 

$$\operatorname{krank}\{\boldsymbol{W}_{\ell}^{\mathsf{T}}\} \geq 2, \quad \operatorname{krank}\{\boldsymbol{W}_{\ell+1}\} \geq 1,$$

where  $\operatorname{krank}\{\boldsymbol{W}_{\ell+1}\} \geq 1$  simply means that  $\boldsymbol{W}_{\ell+1}$  does not have zero columns.

# 4.2 Shallow hPNNs and tensor decompositions

An order-s tensor  $\mathcal{T} \in \mathbb{R}^{m_1 \times \cdots \times m_s}$  is an s-way multidimensional array (more details are provided in Appendix A and more background on tensors can be found in [14–16]). It is said to have a CPD of rank d if it admits a minimal decomposition into d rank-1 terms  $\mathcal{T} = \sum_{j=1}^d a_{1,j} \otimes \cdots \otimes a_{s,j}$  for  $a_{i,j} \in \mathbb{R}^{m_i}$ , with  $\otimes$  being the outer product. The CPD is also written compactly as  $\mathcal{T} = [A_1, A_2, \cdots, A_s]$  for matrices  $A_i = [a_{i,1}, \cdots, a_{i,d}] \in \mathbb{R}^{m_i \times d}$ .  $\mathcal{T}$  is said to be (partially) symmetric if it is invariant to any permutation of (a subset) of its indices [79]. Concretely, if  $\mathcal{T}$  is partially symmetric on dimensions  $i \in \{2, \ldots, s\}$ , its CPD is also partially symmetric with matrices  $A_i$ ,  $i \geq 2$  satisfying  $A_2 = A_3 = \cdots = A_s$ . Our main proofs strongly rely on results of [7] on the connection between hPNN and tensors decomposition in the shallow (i.e., 2-layer) case (see also [79]).

**Proposition 34.** There is a one-to-one mapping between partially symmetric tensors  $\mathcal{F} \in \mathbb{R}^{n \times m \times \cdots \times m}$  and polynomial vectors  $\mathbf{f} \in (\mathscr{P}_{m,r})^{\times n}$ , which can be written as

$$\mathcal{F} \mapsto f(x) = F^{(1)}x^{\otimes r},$$

with  $F^{(1)} \in \mathbb{R}^{n \times m^r}$  the first unfolding of  $\mathcal{F}$ . Under this mapping, the partially symmetric CPD

$$\boldsymbol{\mathcal{F}} = [\![\boldsymbol{W}_2, \boldsymbol{W}_1^\mathsf{T}, \cdots, \boldsymbol{W}_1^\mathsf{T}]\!]$$
 (6)

is mapped to hPNN  $\mathbf{W}_2\rho_r(\mathbf{W}_1\mathbf{x})$ . Thus, uniqueness of hPNN $_{(m,d,n),r}[(\mathbf{W}_1,\mathbf{W}_2)]$  is equivalent to uniqueness of the partially symmetric CPD of  $\mathcal{F}$ .

Thanks to the link with the partially symmetric CPD, we prove the following Kruskal-based sufficient condition for uniqueness (which is a counterpart of Proposition 33).

**Proposition 35.** Let  $p_w(x) = W_2 \rho_{r_1}(W_1 x)$  be a 2-layer hPNN with layer sizes (m, d, n) satisfying  $m, d \geq 2, n \geq 1$ . Assume that  $r \geq 2$ , krank $\{W_2\} \geq 1$ , krank $\{W_1^T\} \geq 2$  and that:

$$r \geq \frac{2d - \operatorname{krank}\{\boldsymbol{W}_2\}}{\operatorname{krank}\{\boldsymbol{W}_1^\mathsf{T}\} - 1} \,,$$

then the 2-layer hPNN  $p_w(x)$  is unique (or equivalently, the CPD of  ${\mathcal F}$  in (6) is unique).

**Remark 36.** For 2-layer hPNNs (L=2), when the activation degree r is high enough Proposition 33 gives both necessary and sufficient conditions for uniqueness due to Proposition 35.

**Remark 37.** Proposition 35 forms the basis of the proof of Proposition 12, which comes from the fact that the Kruskal rank of a generic matrix is equal to its smallest dimension.

**Remark 38.** Proposition 35 is based on basic (Kruskal) uniqueness conditions [80–82]. As mentioned in Remark 15, by using more powerful results on generic uniqueness [83, 84], we can obtain better bounds for identifiability of 2-layer PNNs. For example, for "bottleneck" architectures (as in Corollary 19), the results of [83, Thm 1.11-12] imply that for degrees  $r_{\ell} = 2$ , identifiability holds for decoder layer sizes satisfying a weaker condition  $d_{\ell} \leq \frac{(d_b-1)d_b}{2}$  (instead of  $\frac{d_{\ell}}{r_{\ell}} \leq d_b-1$ ).

## 4.3 Proof of the main result

The proof of Theorem 11 proceeds by induction over the layers  $\ell=1,\ldots,L$ . The key idea is based on a procedure that allows us to prove finite identifiability of the L-th layer given the assumption that the previous layers are identifiable. For this, we introduce a map  $\psi[{\boldsymbol q},{\boldsymbol W}_L]:={\boldsymbol W}_L\rho_{r_{L-1}}({\boldsymbol q}(x_1,\ldots,x_{d_0})),$  where  ${\boldsymbol q}$  is the vector polynomial of degree  $R=r_1\cdots r_{L-2}$ , representing the output of the (L-1)-th linear layer. Then the L-layer hPNN is a composition:

$$\mathrm{hPNN}_{\boldsymbol{r}}[\boldsymbol{\theta},\boldsymbol{W}_L] = \psi[\mathrm{hPNN}_{(r_1,\dots,r_{L-2})}[\boldsymbol{\theta}],\boldsymbol{W}_L], \quad \text{for } \boldsymbol{\theta} = (\boldsymbol{W}_1,\dots,\boldsymbol{W}_{L-1}).$$

To obtain finite identifiability, we look at the Jacobian of the composite map. The key to this recursion is to show that the Jacobian of  $\psi$  with respect to the input polynomial vector and  ${\pmb W}_L$ 

$$J_{\psi}(\boldsymbol{q}, \boldsymbol{W}_{L}) = \begin{bmatrix} J_{\psi}^{(\boldsymbol{q})} & J_{\psi}^{(\boldsymbol{W}_{L})} \end{bmatrix},$$

is of maximal possible rank. For this, we construct a "certificate" of finite identifiability  $q_0$  realized by  $\text{hPNN}_{(r_1,\dots,r_{L-2})}[\theta]$ , but of simpler structure which inherits identifiability of a shallow hPNN.

**Remark 39.** For  $d_L = 1$ , maximality of the rank for  $J_{\psi}^{(\mathbf{q})}$  is closely related to nondefectivity of the variety of sums of powers of forms, which is often proved by establishing Hilbert genericity of an ideal generated by the elements of  $\mathbf{q}$  (a question raised in Fröberg conjecture, see e.g., [85]).

A key limitation of our techniques is that they only allow for establishing finite identifiability for deep PNNs. There exist recent results linking finite and global identifiability, [75, 86] but only for additive decompositions (shallow case). We state, however, the following conjecture.

**Conjecture 40.** Under the assumptions of Theorem 11, the L-layer hPNN is globally identifiable.

Note that the conjecture may be valid only for global identifiability (i.e., for a generic choice of parameters) and not for uniqueness, since it is not true that the composition of unique shallow hPNNs yield a unique deep hPNNs, as shown by the following example.

**Example 41.** Consider two polynomials:  $p(x_1, x_2) = [(x_1^2 + x_2^2)^2 \quad (x_1^2 - x_2^2)^2]^T$ . We see that this polynomial vector admits two different representations

$$p(x) = I \rho_2(W_2 \rho_2(Ix)) = W_3 \rho_2 \left(\frac{1}{2}W_2 \rho_2(W_2x)\right),$$

with

$$\boldsymbol{W}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \boldsymbol{W}_3 = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix},$$

which are not equivalent. However, each 2-layer subnetwork is unique (see Example 7).

# 5 Discussion

In this paper, we presented a comprehensive analysis of the identifiability of deep feedforward PNNs by using their connections to tensor decompositions. Our main result is the *localization of identifiability*, showing that deep PNNs are finitely identifiable if every 2-layer subnetwork is also finitely identifiable for a subset of their inputs. Our results can be also useful for compression (pruning) neural networks as they give an indication about the architectures that are not reducible. An important perspective is also to understand when two different identifiable PNN architectures can represent the same function, as the identifiable representations can potentially occur for different non-compatible formats (e.g., a PNN in format d = (2, 4, 4, 2) could be potentially pruned to two different identifiable representations, say, d = (2, 3, 4, 2) and d = (2, 4, 3, 2).

While our results focus on the case of monomial activations, we believe that this approach can be extended for establishing theoretical guarantees for other types of architectures and activation functions. In fact, the monomial case constitutes as a key first step in addressing general polynomial activations (see, e.g., [45]) which, in turn, can approximate most commonly used activations on compact sets. Moreover, the close connection between PNNs and partially symmetric tensor decompositions (which benefit from efficient computational algorithms based on linear algebra [87]) can also serve as support for the development of computational algorithms based on tensor decompositions for training deep PNNs. In fact, tensor decompositions have been combined with the method of moments to learn small NN architectures (see, e.g., [52, 88]), extending such approaches for training deep PNNs with finite datasets is an important direction for future work.

# Acknowledgments

This work was supported in part by the French National Research Agency (ANR) under grants ANR-23-CE23-0024, ANR-23-CE94-0001, by the PEPR project CAUSALI-T-AI, and by the National Science Foundation, under grant NSF 2316420.

## References

- [1] Grigorios G Chrysos, Stylianos Moschoglou, Giorgos Bouritsas, Yannis Panagakis, Jiankang Deng, and Stefanos Zafeiriou. P-nets: Deep polynomial neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7325–7335, 2020.
- [2] Grigorios G Chrysos, Markos Georgopoulos, Jiankang Deng, Jean Kossaifi, Yannis Panagakis, and Anima Anandkumar. Augmenting deep classifiers with polynomial neural networks. In *European Conference on Computer Vision*, pages 692–716. Springer, 2022.
- [3] Mohsen Yavartanoo, Shih-Hsuan Hung, Reyhaneh Neshatavar, Yue Zhang, and Kyoung Mu Lee. Polynet: Polynomial neural network for 3D shape recognition with polyshape representation. In *International conference on 3D vision (3DV)*, pages 1014–1023. IEEE, 2021.
- [4] Guandao Yang, Sagie Benaim, Varun Jampani, Kyle Genova, Jonathan T. Barron, Thomas Funkhouser, Bharath Hariharan, and Serge Belongie. Polynomial neural fields for subband decomposition and manipulation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=juE5ErmZB61.
- [5] Jie Bu and Anuj Karpatne. Quadratic residual networks: A new class of neural networks for solving forward and inverse problems in physics involving PDEs. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 675–683. SIAM, 2021.
- [6] Sarat Chandra Nayak and Bijan Bihari Misra. Estimating stock closing indices using a GA-weighted condensed polynomial neural network. *Financial Innovation*, 4(1):21, 2018.
- [7] Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. *Advances in neural information processing systems*, 32, 2019.
- [8] Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. On the geometry and optimization of polynomial convolutional networks. AISTATS 2025, 2025. arXiv preprint arXiv:2410.00722.
- [9] Nathan W Henry, Giovanni Luca Marchetti, and Kathlén Kohn. Geometry of lightning self-attention: Identifiability and dimension. *ICLR* 2025, 2025. arXiv preprint arXiv:2408.17221.
- [10] Kaie Kubjas, Jiayi Li, and Maximilian Wiesmann. Geometry of polynomial neural networks. *Algebraic Statistics*, 15(2):295–328, 2024. arXiv:2402.00949.
- [11] Bella Finkel, Jose Israel Rodriguez, Chenxi Wu, and Thomas Yahl. Activation degree thresholds and expressiveness of polynomial neural networks. *Algebraic Statistics*, 16(2):113–130, 2025.
- [12] Samuele Pollaci. Spurious valleys and clustering behavior of neural networks. In *International Conference on Machine Learning*, pages 28079–28099. PMLR, 2023.
- [13] Yossi Arjevani, Joan Bruna, Joe Kileel, Elzbieta Polak, and Matthew Trager. Geometry and optimization of shallow polynomial networks. *arXiv preprint arXiv:2501.06074*, 2025.
- [14] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51 (3):455–500, 2009.
- [15] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on signal processing*, 65(13):3551–3582, 2017.

- [16] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, Danilo P Mandic, et al. Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations and Trends® in Machine Learning*, 9(4-5):249–429, 2016.
- [17] Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Conference on Learning Theory*, pages 742–778. PMLR, 2014.
- [18] Ricardo Borsoi, Konstantin Usevich, and Marianne Clausel. Low-rank tensor decompositions for the theory of neural networks. *arXiv* preprint arXiv:2508.18408, 2025.
- [19] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15: 2773–2832, 2014.
- [20] Héctor J Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural networks*, 5(4):589–593, 1992.
- [21] Francesca Albertini and Eduardo D Sontag. For neural networks, function determines form. *Neural networks*, 6(7):975–990, 1993.
- [22] Francesca Albertini, Eduardo D Sontag, and Vincent Maillot. Uniqueness of weights for neural networks. *Artificial Neural Networks for Speech and Vision*, pages 115–125, 1993.
- [23] Henning Petzka, Martin Trimmel, and Cristian Sminchisescu. Notes on the symmetries of 2-layer ReLU-networks. In *Proceedings of the northern lights deep learning workshop*, volume 1, pages 1–6, 2020.
- [24] Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3):507–555, 1994.
- [25] Verner Vlačić and Helmut Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021.
- [26] Verner Vlačić and Helmut Bölcskei. Neural network identifiability for a family of sigmoidal nonlinearities. *Constructive Approximation*, 55(1):173–224, 2022.
- [27] Flavio Martinelli, Berfin Şimşek, Wulfram Gerstner, and Johanni Brea. Expand-and-cluster: parameter recovery of neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 34895–34919, 2024.
- [28] David Rolnick and Konrad Kording. Reverse-engineering deep ReLU networks. In *International Conference on Machine Learning*, pages 8178–8187. PMLR, 2020.
- [29] Phuong Bui Thi Mai and Christoph Lampert. Functional vs. parametric equivalence of ReLU networks. In 8th International Conference on Learning Representations, 2020.
- [30] Pierre Stock and Rémi Gribonval. An embedding of ReLU networks and an analysis of their identifiability. *Constructive Approximation*, pages 1–47, 2022.
- [31] Joachim Bona-Pellissier, François Malgouyres, and François Bachoc. Local identifiability of deep ReLU neural networks: the theory. *Advances in neural information processing systems*, 35:27549–27562, 2022.
- [32] Joachim Bona-Pellissier, François Bachoc, and François Malgouyres. Parameter identifiability of a deep feedforward ReLU neural network. *Machine Learning*, 112(11):4431–4493, 2023.
- [33] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2021.
- [34] Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, pages 6912–6939. PMLR, 2023.

- [35] Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. Advances in Neural Information Processing Systems, 35:11893–11905, 2022.
- [36] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International conference on machine learning*, pages 4114– 4124. PMLR, 2019.
- [37] Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PUpZXvNqmb.
- [38] Akira Ito, Masanori Yamada, and Atsutoshi Kumagai. Linear mode connectivity between multiple models modulo permutation symmetries. In Forty-second International Conference on Machine Learning, 2025.
- [39] Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.
- [40] Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge university press, 2009.
- [41] Giovanni Luca Marchetti, Vahid Shahverdi, Stefano Mereta, Matthew Trager, and Kathlén Kohn. An invitation to neuroalgebraic geometry. arXiv preprint arXiv:2501.18915, 2025.
- [42] Abbas Kazemipour, Brett W Larsen, and Shaul Druckmann. Avoiding spurious local minima in deep quadratic networks. *arXiv preprint arXiv:2001.00098*, 2019.
- [43] Moulik Choraria, Leello Tadesse Dadi, Grigorios Chrysos, Julien Mairal, and Volkan Cevher. The spectral bias of polynomial neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=P7FLfMLTSEX.
- [44] Pierre Comon, Yang Qi, and Konstantin Usevich. Identifiability of an X-rank decomposition of polynomial maps. SIAM Journal on Applied Algebra and Geometry, 1(1):388–414, 2017. doi: 10.1137/16M1108388.
- [45] Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. Learning on a razor's edge: the singularity bias of polynomial neural networks. *arXiv preprint arXiv:2505.11846*, 2025.
- [46] Steffen Rendle. Factorization machines. In *IEEE International conference on data mining*, pages 995–1000. IEEE, 2010.
- [47] Mathieu Blondel, Masakazu Ishihata, Akinori Fujino, and Naonori Ueda. Polynomial networks and factorization machines: New insights and efficient training algorithms. In *International Conference on Machine Learning*, pages 850–858. PMLR, 2016.
- [48] Mathieu Blondel, Vlad Niculae, Takuma Otsuka, and Naonori Ueda. Multi-output polynomial networks and factorization machines. Advances in Neural Information Processing Systems, 30, 2017.
- [49] Li-Ping Liu, Ruiyuan Gu, and Xiaozhe Hu. Ladder polynomial neural networks. *arXiv preprint arXiv:2106.13834*, 2021.
- [50] Feng-Lei Fan, Mengzhou Li, Fei Wang, Rongjie Lai, and Ge Wang. On expressivity and trainability of quadratic networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [51] Zhijian Zhuo, Ya Wang, Yutao Zeng, Xiaoqing Li, Xun Zhou, and Jinwen Ma. Polynomial composition activations: Unleashing the dynamics of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CbpWPbYHuv.

- [52] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. arXiv preprint arXiv:1506.08473, 2015.
- [53] Massimo Fornasier, Jan Vybíral, and Ingrid Daubechies. Robust and resource efficient identification of shallow neural networks by fewest samples. *Information and Inference: A Journal of the IMA*, 10(2):625–695, 2021.
- [54] Christian Fiedler, Massimo Fornasier, Timo Klock, and Michael Rauchensteiner. Stable recovery of entangled weights: Towards robust identification of deep neural networks from minimal samples. Applied and Computational Harmonic Analysis, 62:123–172, 2023.
- [55] Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.
- [56] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.
- [57] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36:48603–48638, 2023.
- [58] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.
- [59] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.
- [60] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.
- [61] V Lebedev, Y Ganin, M Rakhuba, I Oseledets, and V Lempitsky. Speeding-up convolutional neural networks using fine-tuned CP-decomposition. In Proc. 3rd International Conference on Learning Representations (ICLR), 2015.
- [62] Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Adv. Neur. Inf. Proc. Syst.*, 28, 2015.
- [63] Xingyi Liu and Keshab K Parhi. Tensor decomposition for model reduction in neural networks: A review. *IEEE Circuits and Systems Magazine*, 23(2):8–28, 2023.
- [64] Anh-Huy Phan, Konstantin Sobolev, Konstantin Sozykin, Dmitry Ermilov, Julia Gusak, Petr Tichavskỳ, Valeriy Glukhov, Ivan Oseledets, and Andrzej Cichocki. Stable low-rank tensor decomposition for compression of convolutional neural network. In *Proc. 16th European Conference on Computer Vision (ECCV)*, pages 522–539, Glasgow, UK, 2020. Springer.
- [65] Emanuele Zangrando, Steffen Schotthöfer, Jonas Kusch, Gianluca Ceruti, and Francesco Tudisco. Geometry-aware training of factorized layers in tensor Tucker format. *Proceedings*, Adv. Neur. Inf. Proc. Syst., 2024.
- [66] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pages 698–728. PMLR, 2016.
- [67] Maude Lizaire, Michael Rizvi-Martel, Marawan Gamal, and Guillaume Rabusseau. A tensor decomposition perspective on second-order RNNs. In *Forty-first International Conference on Machine Learning*, 2024.
- [68] Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. In *ICLR*, 2018.

- [69] Anuj Mahajan, Mikayel Samvelyan, Lei Mao, Viktor Makoviychuk, Animesh Garg, Jean Kossaifi, Shimon Whiteson, Yuke Zhu, and Animashree Anandkumar. Tesseract: Tensorised actors for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 7301–7312. PMLR, 2021.
- [70] Sergio Rozada, Santiago Paternain, and Antonio G Marques. Tensor and matrix low-rank valuefunction approximation in reinforcement learning. *IEEE Transactions on Signal Processing*, 2024.
- [71] Marco Mondelli and Andrea Montanari. On the connection between learning two-layer neural networks and tensor decomposition. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1051–1060. PMLR, 2019.
- [72] Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. In *International Conference on Learning Representations*, 2019.
- [73] Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Efficient algorithms for learning depth-2 neural networks with general ReLU activations. Adv. Neur. Inf. Proc. Syst., 34:13485– 13496, 2021.
- [74] François Malgouyres and Joseph Landsberg. Multilinear compressive sensing and an application to convolutional linear networks. SIAM Journal on Mathematics of Data Science, 1(3):446–475, 2019.
- [75] Alex Casarotti and Massimiliano Mella. From non-defectivity to identifiability. *Journal of the European Mathematical Society*, 25(3):913–931, 2022.
- [76] Joseph M Landsberg. Tensors: Geometry and applications, volume 128. American Mathematical Soc., 2012.
- [77] Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *International Conference on Machine Learning*, pages 3560–3569. PMLR, 2019.
- [78] Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B Grosse. Regularized linear autoencoders recover the principal components, eventually. *Advances in Neural Information Processing Systems*, 33:6971–6981, 2020.
- [79] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. SIAM Journal on Matrix Analysis and Applications, 30(3):1254–1279, 2008.
- [80] Nicholas D Sidiropoulos and Xiangqian Liu. Identifiability results for blind beamforming in incoherent multipath with small delay spread. *IEEE Transactions on Signal Processing*, 49(1): 228–236, 2001.
- [81] Ignat Domanov and Lieven De Lathauwer. On the uniqueness of the canonical polyadic decomposition of third-order tensors—Part II: Uniqueness of the overall decomposition. *SIAM Journal on Matrix Analysis and Applications*, 34(3):876–903, 2013.
- [82] Nicholas D. Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, 14(3):229–239, 2000.
- [83] Ignat Domanov and Lieven De Lathauwer. Generic uniqueness conditions for the canonical polyadic decomposition and INDSCAL. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1567–1589, 2015.
- [84] Hirotachi Abo and Maria Chiara Brambilla. On the dimensions of secant varieties of Segre-Veronese varieties. *Annali di Matematica Pura ed Applicata*, 192(1):61–92, 2013.
- [85] Ralf Fröberg, Samuel Lundqvist, Alessandro Oneto, and Boris Shapiro. Algebraic stories from one and from the other pockets. *Arnold Mathematical Journal*, 4(2):137–160, 2018.
- [86] Alex Massarenti and Massimiliano Mella. Bronowski's conjecture and the identifiability of projective varieties. *Duke Mathematical Journal*, 173(17):3293–3316, 2024.

- [87] Kim Batselier and Ngai Wong. Symmetric tensor decomposition by an iterative eigendecomposition algorithm. *Journal of Computational and Applied Mathematics*, 308:69–82, 2016.
- [88] Samet Oymak and Mahdi Soltanolkotabi. Learning a deep convolutional neural network via tensor decomposition. *Information and Inference: A Journal of the IMA*, 10(3):1031–1071, 2021.

# A Background on tensors and results for shallow networks

In this appendix, we first present a background on tensors and tensor decompositions and some technical results. We start with basic definitions about tensors and the CP decomposition (with especial emphasis to the symmetric and partially symmetric cases). Then, we introduce Kruskal-based uniqueness conditions and some related technical lemmas. Finally, we demonstrate the link between hPNNs and the partially symmetric CPD and based on this connection, we derive sufficient uniqueness conditions for 2-layer hPNNs. We present both necessary and sufficient uniqueness conditions for the 2-layer case based on Kruskal's conditions and the uniqueness of tensors decompositions.

**Results from the main paper**: Lemmas 30 and 31, Propositions 33, 34 and 35.

# A.1 Basics on tensors and tensor decompositions

**Notation.** The order of a tensor is the number of dimensions, also known as ways or modes. Vectors (tensors of order one) are denoted by boldface lowercase letters, e.g., a. Matrices (tensors of order two) are denoted by boldface capital letters, e.g., A. Higher-order tensors (order three or higher) are denoted by boldface Euler script letters, e.g.,  $\mathcal{X}$ .

**Unfolding of tensors.** The *p*-th unfolding (also called mode-*p* unfolding) of a tensor of order *s*,  $\mathcal{T} \in \mathbb{R}^{m_1 \times \cdots \times m_s}$  is the matrix  $\mathbf{T}^{(p)}$  of size  $m_r \times (m_1 m_2 \cdots m_{r-1} m_{r+1} \cdots m_s)$  defined as

$$\left[ m{T}^{(p)} 
ight]_{i_r,j} = m{\mathcal{T}}_{i_1,\dots,i_r,\dots,i_s}, \ ext{where} \ j = 1 + \sum_{\substack{n=1 \ n 
eq r}}^s (i_n - 1) \prod_{\substack{\ell = 1 \ \ell 
eq r}}^{n-1} m_\ell \,.$$

We give an example of unfolding extracted from [14]. Let the frontal slices of  $\mathcal{X} \in \mathbb{R}^{3 \times 4 \times 2}$  be

$$\boldsymbol{X}_1 = \begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix}, \ \boldsymbol{X}_2 = \begin{pmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{pmatrix}.$$

Then the three mode-n unfoldings of  $\mathcal{X}$  are

$$\boldsymbol{X}^{(1)} = \begin{pmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{pmatrix}$$

$$\boldsymbol{X}^{(2)} = \begin{pmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{pmatrix}$$

$$\boldsymbol{X}^{(3)} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & \cdots & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & \cdots & 22 & 23 & 24 \end{pmatrix}$$

Symmetric and partially symmetric tensors. A tensor of order  $s, \mathcal{T} \in \mathbb{R}^{m_1 \times \cdots \times m_s}$  is said to be symmetric if  $m_1 = \cdots = m_s$  and for every permutation  $\sigma$  of  $\{1, \ldots, s\}$ :

$$\mathcal{T}_{i_1,i_2,\cdots,i_s} = \mathcal{T}_{i_{\sigma(1)},i_{\sigma(2)},\ldots,i_{\sigma(s)}}.$$

The tensor  $\mathcal{T} \in \mathbb{R}^{m_1 \times \cdots \times m_s}$  is said to be *partially symmetric* along the modes  $(r+1,\ldots,s)$  for r < s if  $m_{r+1} = \cdots = m_s$  and for every permutation  $\sigma$  of  $\{r+1,\ldots,s\}$ 

$${\cal T}_{i_1,i_2,...,i_r,i_{r+1},\cdots,i_s} = {\cal T}_{i_1,...,i_r,i_{\sigma(r+1)},...,i_{\sigma(s)}}.$$

**Mode products.** The r-mode (matrix) product of a tensor  $\mathcal{T} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_s}$  with a matrix  $A \in \mathbb{R}^{J \times m_r}$  is denoted by  $\mathcal{T} \bullet_r A$  and is of size  $m_1 \times \cdots \times m_{r-1} \times J \times m_{r+1} \times \cdots \times m_s$ . It is defined elementwise, as

$$egin{aligned} \left[oldsymbol{\mathcal{T}}ullet_roldsymbol{A}
ight]_{i_1,...,i_{r-1},j,i_{r+1},...,i_s} = \sum_{i_r=1}^{m_r}oldsymbol{\mathcal{T}}_{i_1,...,i_s}oldsymbol{A}_{j,i_r}\,. \end{aligned}$$

**Minimal rank-**R **decomposition.** The canonical polyadic decomposition (CPD) of a tensor T is the decomposition of a tensor as a sum of R rank-1 tensors where R is minimal [14, 15], that is

$$\mathcal{T} = \sum_{i=1}^R \boldsymbol{a}_i^{(1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(s)},$$

where, for each  $p \in \{1, \cdots, s\}$ ,  $a_i^{(p)} \in \mathbb{R}^{m_p}$ , and  $\otimes$  denotes the outer product operation. Alternatively, we denote the CPD by

$$\mathcal{T} = \llbracket \boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \dots, \boldsymbol{A}^{(s)} 
rbracket,$$

where 
$$oldsymbol{A}^{(p)} = ig[oldsymbol{a}_1^{(p)} \cdots oldsymbol{a}_R^{(p)}ig] \in \mathbb{R}^{m_p imes R}.$$

When  $\mathcal{T}$  is partially symmetric along the modes  $(p+1,\ldots,s)$ , for p< s, its CPD satisfies  $\mathbf{A}^{(p+1)}=\mathbf{A}^{(p+2)}=\cdots=\mathbf{A}^{(s)}$ . The case of fully symmetric tensors (i.e., tensors which are symmetric along all their dimensions) deserves special attention [79]. The CPD of a fully symmetric tensor  $\mathcal{T}\in\mathbb{R}^{m\times m\times \cdots\times m}$  is defined as

$$\mathcal{T} = \sum_{i=1}^{R} u_i \, \boldsymbol{a}_i \otimes \cdots \otimes \boldsymbol{a}_i,$$

where  $u_i \in \mathbb{R}$  are real-valued coefficients. With a slight abuse of notation, we represent it compactly using the same notation as an order-(n+1) tensor of size  $1 \times m \times \cdots \times m$ , as

$$\mathcal{T} = \llbracket \boldsymbol{u}, \boldsymbol{A}, \cdots, \boldsymbol{A} 
Vert,$$

where  $u \in \mathbb{R}^{1 \times m}$  is a  $1 \times m$  matrix (i.e., a row vector) containing the coefficients  $u_i$ , that is,  $u_i = u_i$ ,  $i = 1, \dots, R$ .

# A.1.1 Kruskal-based conditions

To obtain sufficient conditions for the uniqueness of 2-layer and deep hPNNs, we first need some preliminary technical results about tensor decompositions.

Let  $\odot$  denote the column-wise Khatri-Rao product and denote by  ${m A}^{\odot k}$  the k-th Khatri-Rao power:

$$oldsymbol{A}^{\odot k} = \underbrace{oldsymbol{A} \odot \cdots \odot oldsymbol{A}}_{k ext{ times}}.$$

We recall the following well known lemma:

**Lemma A.1.** Let  $A \in \mathbb{R}^{I \times R}$ , then the Kruskal rank of its k-th Khatri-Rao power satisfies  $\operatorname{krank}\{A^{\odot k}\} > \min(R, k \operatorname{krank}\{A\} - k + 1)$ .

The proof of Lemma A.1 can be found in [80, Lemma 1] or in [81, Corollary 1.18].

We will also need another two well-known lemmas.

Lemma A.2. Let A full column rank. Then

$$krank\{AB\} = krank\{B\}$$

for any compatible matrix B.

*Proof.* Since A has maximal rank, for any columns  $B_{:,j_1},\ldots,B_{:,j_k}$  of B, one has  $\dim \operatorname{span}((AB)_{:,j_1},\ldots,(AB)_{:,j_k})=\dim A \cdot \operatorname{span}(B_{:,j_1},\ldots,B_{:,j_k})=\dim \operatorname{span}(B_{:,j_1},\ldots,B_{:,j_k})$ . By definition of the Kruskal rank,  $\operatorname{krank}\{AB\}=\operatorname{krank}\{B\}$ .

**Lemma A.3** (Kruskal's theorem, s-way version [82], Thm. 3). Let  $\mathcal{T} = [\![ \boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \cdots, \boldsymbol{A}^{(s)} ]\!]$  be a tensor with CP rank R and  $\boldsymbol{A}^{(i)} \in \mathbb{R}^{m_i \times R}$ , such that

$$\sum_{i=1}^{s} \operatorname{krank}\{A^{(i)}\} \ge 2R + (s-1). \tag{7}$$

Then the CP decomposition of  $\mathcal{T}$  is unique up to permutation and scaling ambiguities, that is, for any alternative CPD  $\mathcal{T} = [\![\widetilde{\boldsymbol{A}}^{(1)}, \widetilde{\boldsymbol{A}}^{(2)}, \cdots, \widetilde{\boldsymbol{A}}^{(s)}]\!]$ , there exist a permutation matrix  $\boldsymbol{\Pi}$  and invertible diagonal matrices  $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_s$  such that

$$\widetilde{\boldsymbol{A}}^{(i)} = \boldsymbol{A}^{(i)} \boldsymbol{\Pi} \boldsymbol{\Lambda}_1$$

for  $i = 1, \ldots, s$ .

## A.1.2 Link between hPNNs and partially symmetric tensors

Recall that  $\mathscr{P}_{m,r}$  denotes the space of m-variate polynomials of degree  $\leq r$ . The following proposition, originally presented in Section 4 of the main body of the paper, formalizes the link between polynomial vectors and partially symmetric tensors.

**Proposition 34.** There is a one-to-one mapping between partially symmetric tensors  $\mathcal{F} \in \mathbb{R}^{n \times m \times \dots \times m}$  and polynomial vectors  $\mathbf{f} \in (\mathscr{P}_{m,r})^{\times n}$ , which can be written as

$$\mathcal{F} \mapsto f(x) = F^{(1)}x^{\otimes r},$$

with  $F^{(1)} \in \mathbb{R}^{n \times m^r}$  the first unfolding of  $\mathcal{F}$ . Under this mapping, the partially symmetric CPD

$$\mathcal{F} = \llbracket \boldsymbol{W}_2, \boldsymbol{W}_1^\mathsf{T}, \cdots, \boldsymbol{W}_1^\mathsf{T} 
brace$$

is mapped to hPNN  $W_2\rho_r(W_1x)$ . Thus, uniqueness of  $\mathrm{hPNN}_{(m,d,n),(r)}[(W_1,W_2)]$  is equivalent to uniqueness of the partially symmetric CPD of  $\mathcal{F}$ .

*Proof.* We distinguish the two cases, n=1 and  $n\geq 2$ . We begin the proof by the more general case  $n\geq 2$ .

Case  $n \geq 2$ . Denoting by  $u_i \in \mathbb{R}^n$  the *i*-th column of  $W_2$  and  $v_i \in \mathbb{R}^m$  the *i*-th row of  $W_1$ , the relationship between the 2-layer hPNN and tensor  $\mathcal{F}$  can be written explicitly as

$$egin{aligned} oldsymbol{f}(oldsymbol{x}) &= oldsymbol{W}_2 
ho_r(oldsymbol{W}_1 oldsymbol{x}) \ &= \sum_{i=1}^d oldsymbol{u}_i (oldsymbol{v}_i^\mathsf{T} oldsymbol{x})^\mathsf{T} oldsymbol{x}^{\otimes r} \ &= oldsymbol{W}_2 ig(oldsymbol{W}_1^\mathsf{T} \odot \cdots \odot oldsymbol{W}_1^\mathsf{T}ig)^\mathsf{T} oldsymbol{x}^{\otimes r} \,, \ &= oldsymbol{W}_2 ig(oldsymbol{W}_1^\mathsf{T} \odot \cdots \odot oldsymbol{W}_1^\mathsf{T}ig)^\mathsf{T} oldsymbol{x}^{\otimes r} \,, \end{aligned}$$

where  $\odot$  denotes the Khatri-Rao product. The equivalence of the last expression and the first unfolding of the order-(r+1) tensor  $\mathcal{F}$  can be found in [14].

The special case n=1. When n=1, the columns of  $\boldsymbol{W}_2 \in \mathbb{R}^{1 \times d}$  are scalars values  $u_i \in \mathbb{R}$ ,  $i=1,\ldots,d$ . In this case,  $(\boldsymbol{W}_1^\mathsf{T} \odot \cdots \odot \boldsymbol{W}_1^\mathsf{T}) \boldsymbol{W}_2^\mathsf{T}$  becomes equivalent to the vectorization of  $\boldsymbol{\mathcal{F}}$ , which is a fully symmetric tensor of order r with factors  $\boldsymbol{W}_1^\mathsf{T}$  and coefficients  $[\boldsymbol{W}_2]_{1,i}, i=1,\ldots,d$ .  $\square$ 

## A.1.3 Technical lemmas

In this subsection we prove the key lemmas stated in Section 4 (Lemma 30 and Lemma 31). These results give necessary conditions for the uniqueness of an hPNN in terms of the minimality of an unique architectures and the independence (non-redundancy) of its internal representations, as well as a connection between the uniqueness of two 2-layer hPNNs based on the concision of tensors. They will be used in the proof of the localization theorem.

**Lemma 30.** Let  $p = \text{hPNN}_r[w]$  be an hPNN of format  $(d_0, \dots, d_\ell, \dots, d_L)$ . Then for any  $\ell$  there exists an infinite number of representations of hPNNs  $p = \text{hPNN}_r[w]$  with architecture  $(d_0, \dots, d_\ell + 1, \dots, d_L)$ . In particular, the augmented hPNN is not unique (or finite-to-one).

Proof of Lemma 30. Let  $(\boldsymbol{W}_0,\cdots,\boldsymbol{W}_L)$  the weight matrices associated with the representation of format  $(d_0,\ldots,d_\ell,\ldots,d_L)$  of the hPNN  $\boldsymbol{p}=\text{hPNN}_{\boldsymbol{r}}[\boldsymbol{w}]$ . By assumptions on the dimensions, the two matrices  $\boldsymbol{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$  and  $\boldsymbol{W}_{\ell+1} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$  read

$$egin{aligned} oldsymbol{W}_{\ell+1} = & [oldsymbol{w}_1 & \cdots & oldsymbol{w}_{d_\ell}] \ oldsymbol{w}_\ell = & [oldsymbol{v}_1 & \cdots & oldsymbol{v}_{d_\ell}]^\mathsf{T} \ , \ ext{where, for each } i, \ oldsymbol{v}_i \in \mathbb{R}^{d_{\ell-1}} \ . \end{aligned}$$

Without loss of generality, let us assume that  $w_i$  are nonzero, and set

$$oldsymbol{\widetilde{W}}_{\ell} = egin{bmatrix} \mathbf{0} & oldsymbol{v}_1 & \cdots & oldsymbol{v}_{d_{\ell}} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{(d_{\ell}+1) imes d_{\ell-1}}$$
 ,

in which we add a row of zeroes to  $W_{\ell}$ . In this case, we can take the following family of matrices defined for any  $u \in \mathbb{R}^{d_{\ell+1}}$ :

$$\widetilde{oldsymbol{W}}_{\ell+1}^{(oldsymbol{u})} = [oldsymbol{u} \quad oldsymbol{w}_1 \quad \cdots \quad oldsymbol{w}_{d_\ell}] \in \mathbb{R}^{d_{\ell+1} imes (d_\ell+1)} \,.$$

Then, we have that for any choice of u and for any z,

$$\widetilde{\boldsymbol{W}}_{\ell+1}^{(\boldsymbol{u})} \rho_{r_\ell}(\widetilde{\boldsymbol{W}}_\ell \boldsymbol{z}) = \boldsymbol{W}_{\ell+1} \rho_{r_\ell}(\boldsymbol{W}_\ell \boldsymbol{z}) \,.$$

The matrices  $\widetilde{W}_{\ell+1}^{(0)}$  and  $\widetilde{W}_{\ell+1}^{(u)}$  for  $u \neq 0$  have a different number of zero columns and cannot be a permutation/rescaling of each other, constituting different representations of the same hPNN p. In fact, every choice of u' that is not collinear to u and  $w_i$ ,  $i = 1, \ldots, d_\ell$  leads to a different non-equivalent representation of p. Thus, we have an infinite number of non-equivalent representations

$$(\boldsymbol{W}_0,\ldots,\boldsymbol{W}_{\ell-1},\widetilde{\boldsymbol{W}}_\ell,\widetilde{\boldsymbol{W}}_{\ell+1}^{(\boldsymbol{u})},\ldots,\boldsymbol{W}_L)$$

of format  $(d_0, \ldots, d_\ell + 1, \ldots, d_L)$  for the hPNN  $\boldsymbol{p} = \text{hPNN}_{\boldsymbol{r}}[\boldsymbol{w}]$ .

Lemma 30 can be seen as a form of minimality or irreducibility of unique hPNNs, as it shows that a unique hPNN does not admit a smaller (i.e., with a lower number of neurons) representation.

**Lemma 31.** For the widths  $\mathbf{d} = (d_0, \dots, d_L)$ , let  $\mathbf{p} = \mathrm{hPNN}_{\mathbf{r}}[\mathbf{w}]$  be a unique L-layers decomposition. Consider the vector output at any  $\ell$ -th internal level  $\ell < L$  after the activations

$$q_{\ell}(\boldsymbol{x}) = \rho_{r_{\ell}} \circ \boldsymbol{W}_{\ell} \circ \cdots \circ \rho_{r_{1}} \circ \boldsymbol{W}_{1}(\boldsymbol{x}).$$

Then the elements  $q_{\ell}(x) = [q_{\ell,1}(x) \cdots q_{\ell,d_{\ell}}(x)]^{\mathsf{T}}$  are linearly independent polynomials.

*Proof of Lemma 31.* By contradiction, suppose that the polynomials  $q_{\ell,1}(\boldsymbol{x}),\dots,q_{\ell,d_{\ell}}(\boldsymbol{x})$  are linearly dependent. Assume without loss of generality that, e.g., the last polynomial  $q_{\ell,d_{\ell}}(\boldsymbol{x})$  can expressed as a linear combination of the others. Then, there exists a matrix  $\boldsymbol{B} \in \mathbb{R}^{d_{\ell} \times (d_{\ell} - 1)}$  so that

$$m{p} = m{W}_L \circ 
ho_{r_{L-1}} \circ \cdots \circ 
ho_{r_{\ell+1}} \circ m{W}_{\ell+1} m{B} egin{bmatrix} q_{\ell,1}(m{x}) \ dots \ q_{\ell,d_{\ell}-1}(m{x}) \end{bmatrix},$$

i.e., the hPNN p admits a representation of size  $d = (d_0, \ldots, d_\ell - 1, \ldots, d_L)$  with parameters  $(W_1, \ldots, W_{\ell+1}B, \ldots, W_L)$ . Therefore, by Lemma 30 its original representation is not unique, which is a contradiction.

# A.1.4 Necessary conditions for uniqueness

Using Lemma 31 and Lemma 30, we can prove the conditions on the Kruskal ranks of weight matrices that are necessary for uniqueness.

**Proposition 33.** As in Lemma 31, let the widths be  $\mathbf{d} = (d_0, \dots, d_L)$ , and  $\mathbf{p} = \mathrm{hPNN}_{\mathbf{r}}[\mathbf{w}]$  have a unique (or finite-to-one) L-layers decomposition. Then we have that for all  $\ell = 1, \dots, L-1$ 

$$\operatorname{krank}\{\boldsymbol{W}_{\ell}^{\mathsf{T}}\} \geq 2, \quad \operatorname{krank}\{\boldsymbol{W}_{\ell+1}\} \geq 1,$$

where  $\operatorname{krank}\{W_{\ell+1}\} \geq 1$  simply means that  $W_{\ell+1}$  does not have zero columns.

Proof of Proposition 33. Suppose that  $\operatorname{krank}\{\boldsymbol{W}_{\ell}^{\mathsf{T}}\}$  < 2. Then we have that at level  $\ell$ , the vector  $q_{\ell}(\boldsymbol{x})$  of internal features defined in (5) contains linearly dependent or zero polynomials, which violates Lemma 31.

Similarly if  $\operatorname{krank}\{W_{\ell+1}\}=0$ , then the neuron corresponding to the zero column can be pruned to obtain a representation with  $(d_\ell-1)$  neurons at the  $\ell$ -th level, which implies loss of uniqueness by Lemma 30 and thus leads to a contradiction.

## A.2 Kruskal-based conditions for the uniqueness of 2-layer networks (L=2)

This subsection provides sufficient conditions for the uniqueness of 2-layer hPNNs. These conditions use Kruskal-based results from tensor decompositions and complement necessary conditions from Appendix A.1.4.

## A.2.1 Sufficient conditions for uniqueness

Now we prove Proposition 35 giving sufficient conditions for uniqueness in the case L=2.

**Proposition 35.** Let  $p_w(x) = W_2 \rho_{r_1}(W_1 x)$  be a 2-layer hPNN with  $W_1 \in \mathbb{R}^{d \times m}$  and  $W_2 \in \mathbb{R}^{n \times d}$  and layer sizes (m, d, n) satisfying  $m, d \geq 2$ ,  $n \geq 1$ . Assume that  $r \geq 2$ ,  $\operatorname{krank}\{W_2\} \geq 1$ ,  $\operatorname{krank}\{W_1^{\mathsf{T}}\} \geq 2$  and that:

$$r \geq \frac{2d - \operatorname{krank}\{\boldsymbol{W}_2\}}{\operatorname{krank}\{\boldsymbol{W}_1^\mathsf{T}\} - 1} \,,$$

then the 2-layer hPNN  $p_w(x)$  is unique (or equivalently, the CPD of  ${\mathcal F}$  in (6) is unique).

*Proof of Proposition 35.* One can apply Proposition 34 to show that the 2-layer hPNN  $p_w(x)$  is in one-to-one correspondence with the order r+1 partially symmetric tensor

$$\mathcal{F} = [ \boldsymbol{W}_2, \boldsymbol{W}_1^\mathsf{T}, \cdots, \boldsymbol{W}_1^\mathsf{T} ], \tag{8}$$

thus, the uniqueness of  $p_w(x)$  is equivalent to that of the CP-decomposition of  $\mathcal F$  in (8). From [82, Theorem 3], the rank-d CP decomposition of  $\mathcal T$  is unique provided that

$$\operatorname{krank}\{\boldsymbol{W}_{2}\} + r \operatorname{krank}\{\boldsymbol{W}_{1}^{\mathsf{T}}\} \geq 2d + r$$
.

By noting that  $krank\{W_1^{\mathsf{T}}\} > 1$  and rearranging the terms, we obtain the desired result.

Note that for the case of  $m \geq 2$  (i.e., hPNNs with at least two outputs), Proposition 35 gives conditions that hold for quadratic activation degrees  $r \geq 2$ . On the other hand, for networks with a single output (i.e., n=1), it requires  $r \geq 3$ .

# A.2.2 Sufficient conditions for identifiability

Equipped with the sufficient conditions for the uniqueness of 2-layer hPNNs obtained in Proposition 35, we can now prove the generic identifiability result stated in Proposition 12.

**Proposition 12.** Let  $m, d \geq 2, n \geq 1$  be the layer widths and  $r \geq 2$  such that

$$r \ge \frac{2d - \min(d, n)}{\min(d, m) - 1}.$$

Then the 2-layer hPNN with architecture ((m, d, n), (r)) is globally identifiable.

*Proof of Proposition 12.* For general matrices  $W_1 \in \mathbb{R}^{d \times m}$  and  $W_2 \in \mathbb{R}^{n \times d}$ , we have

$$krank\{\boldsymbol{W}_{1}^{\mathsf{T}}\} = \min(m, d),$$
$$krank\{\boldsymbol{W}_{2}\} = \min(n, d).$$

Moreover,  $m, d \geq 2, n \geq 1$  implies that generically  $\operatorname{krank}\{\boldsymbol{W}_1^\mathsf{T}\} \geq 2$  and  $\operatorname{krank}\{\boldsymbol{W}_2\} \geq 1$ . This along with (4) means that the assumptions in Proposition 35 are satisfied for all parameters except for a set of Lebesgue measure zero. Thus, the hPNN with architecture ((m,d,n),(r)) is globally identifiable.

# **B** Proof of the localization theorem

This appendix contains the main proofs of the localization theorem (Theorem 11) for deep hPNNs, as well as supporting lemmas and auxiliary technical results. We also provide proofs of the corollaries that specialize this result for several choices of architectures (e.g., pyramidal, bottleneck) and to the activation thresholds, discussed in Section 3.2 of the main paper.

**Results from the main paper**: Theorem 11, Corollaries 16, 19, and 17.

## B.1 Preparatory lemmas - rank of Jacobian of a 2-layer PNN

**Lemma B.1.** Let (m, d, n) and r, so that the 2-layer hPNN with architecture ((m, d, n), r) is finitely identifiable (resp. the partially symmetric rank-d decomposition of  $n \times m \times \cdots \times m$  tensor). Then for general matrices V, W the Jacobian of the map  $p[V, W] = \text{hPNN}_r[(V, W)]$ , given by

$$J_{\boldsymbol{p}}(\boldsymbol{V}, \boldsymbol{W}) = \begin{bmatrix} J_{\boldsymbol{p}}^{(\boldsymbol{V})} & J_{\boldsymbol{p}}^{(\boldsymbol{W})} \end{bmatrix},$$

has maximal possible rank:

$$rank\{J_{p}(\boldsymbol{V}, \boldsymbol{W})\} = (m+n-1)d, \tag{9}$$

and also

$$\operatorname{rank}\{J_{\mathbf{p}}^{(V)}\} = md. \tag{10}$$

*Proof.* The first statement follows from dimension of the neurovariety (that is (m+n-1)d), and the second statement follows from the fact that the subset of pairs (V, W) with W given as

$$oldsymbol{W} = \begin{bmatrix} 1 & \cdots & 1 \\ \overline{oldsymbol{W}} \end{bmatrix}, \quad \overline{oldsymbol{W}} \in \mathbb{R}^{(n-1) \times d}$$

parameterizes an open subset of the neurovariety (i.e., by the scaling ambiguity, almost any pair of  $\widetilde{V}$  and  $\widetilde{W}$  can be reduced to such a form). Therefore, the reduced Jacobian is full column rank:

$$\operatorname{rank}\{\begin{bmatrix}J_{\boldsymbol{p}}^{(\boldsymbol{V})} & J_{\boldsymbol{p}}^{(\overline{\boldsymbol{W}})}\end{bmatrix}\} = md + (n-1)d,$$

which implies the rank condition on  $J_{\mathbf{p}}^{(\mathbf{V})}$ .

**Remark B.2.** The conditions in Lemma B.1 are satisfied, for example, if the Kruskal-based generic uniqueness conditions are satisfied (see Proposition 12):

$$r \ge \frac{2d - \min(d, n)}{\min(d, m) - 1}.$$

To give more intuition we give an example of m=2, n=1, where the inequality reads  $r\geq 2d-1$ . **Example B.3.** Consider bivariate single-output hPNN with  $\mathbf{W}=[1 \ \cdots \ 1]$  and

$$\boldsymbol{V} = \begin{bmatrix} \alpha_1 & \beta_1 \\ \vdots & \vdots \\ \alpha_d & \beta_d \end{bmatrix},$$

so that

$$p[V] = \sum_{j=1}^{d} (\alpha_j x_1 + \beta_j x_2)^r$$
.

Then, the columns of the Jacobian of  $J_p^{(V)}$  are coefficients of polynomials

$$rx_1(\alpha_j x_1 + \beta_j x_2)^{r-1}, rx_2(\alpha_j x_1 + \beta_j x_2)^{r-1},$$

so they can be represented in matrix form as

$$J_p^{(V)} = \begin{bmatrix} r\alpha_1^{r-1} & 0 & \cdots & r\alpha_d^{r-1} & 0 \\ (r-1)\alpha_1^{r-2}\beta_1 & \alpha_1^{r-1} & \cdots & (r-1)\alpha_d^{r-2}\beta_d & \alpha_d^{r-1} \\ (r-2)\alpha_1^{r-3}\beta_1^2 & 2\alpha_1^{r-2}\beta_1 & \cdots & (r-2)\alpha_d^{r-3}\beta_d^2 & 2\alpha_d^{r-2}\beta_d \\ \vdots & & & \vdots \\ (r-s)\alpha_1^{r-s-1}\beta_1^s & s\alpha_1^{r-s}\beta_1^{s-1} & \cdots & (r-s)\alpha_d^{r-s-1}\beta_d^s & s\alpha_d^{r-s}\beta_d^{s-1} \\ \vdots & & & & \vdots \\ \beta_1^{r-1} & (r-1)\alpha_1\beta_1^{r-2} & \cdots & \beta_d^{r-1} & (r-1)\alpha_d\beta_d^{r-2} \\ 0 & r\beta_1^{(r-1)} & \cdots & 0 & r\beta_d^{(r-1)} \end{bmatrix}.$$

This is a confluent Vandermonde matrix and it is known that  $\operatorname{rank}\{J_p^{(V)}\}=2d$  provided  $r\geq 2d-1$  and  $\operatorname{krank}\{V^{\mathsf{T}}\}\geq 2$  (none of pairs  $(\alpha_i,\beta_i)$  and  $(\alpha_\ell,\beta_\ell)$  are collinear).

**Remark B.4** (Explicit form of the Jacobian in the general case). Let (m, d, n), r, V and W be as in Lemma B.1. With some abuse of notation we denote  $v_j \in \mathbb{R}^m$  and  $w_i \in \mathbb{R}^n$ 

$$oldsymbol{V}^\mathsf{T} = [oldsymbol{v}_1 \quad \cdots \quad oldsymbol{v}_d] \,, \quad oldsymbol{W} = [oldsymbol{w}_1 \quad \cdots \quad oldsymbol{w}_d] \,,$$

and let  $z = \begin{bmatrix} z_1 & \cdots & z_m \end{bmatrix}^\mathsf{T}$ . Then the PNN reads

$$p[V, W] = \sum_{j=1}^{d} w_j (v_j^{\mathsf{T}} z)^r.$$
(11)

Therefore, we have that derivatives with respect to the elements of the matrix W can be expressed as

$$\frac{\partial}{\partial W_{i,j}} \boldsymbol{p} = \frac{\partial}{\partial (\boldsymbol{w}_i)_i} \boldsymbol{p} = \boldsymbol{e}_i (\boldsymbol{v}_j^\mathsf{T} \boldsymbol{z})^r$$
 (12)

and, with respect to elements of V, we have

$$\frac{\partial}{\partial V_{j,\ell}} \boldsymbol{p} = \frac{\partial}{\partial (\boldsymbol{v}_j)_{\ell}} \boldsymbol{p} = (rz_{\ell}) \cdot \boldsymbol{w}_j (\boldsymbol{v}_j^{\mathsf{T}} \boldsymbol{z})^{r-1}.$$
(13)

Therefore Lemma B.1 concerns the dimensions of these sets of polynomials.

## **B.2** Structure of composite Jacobian: statement and examples

We can formulate the following proposition:

**Proposition B.5.** Let  $d_0, d, n$  and  $r, R \ge 2$  be fixed. Consider the following map that maps a homogeneous polynomial vector of degree R to homogeneous polynomial vector of degree Rr:

$$\psi: (\mathscr{H}_{d_0,R})^d \times \mathbb{R}^{n \times d} \to (\mathscr{H}_{d_0,Rr})^n$$

$$(\boldsymbol{q}(x_1,\ldots,x_{d_0}),\boldsymbol{W}) \mapsto \psi[\boldsymbol{q},\boldsymbol{W}] := \boldsymbol{W}\rho_r(\boldsymbol{q}(x_1,\ldots,x_{d_0})),$$

and denote the Jacobian with respect to the parameters as

$$J_{\psi}(\boldsymbol{q}, \boldsymbol{W}) = \begin{bmatrix} J_{\psi}^{(\boldsymbol{q})} & J_{\psi}^{(\boldsymbol{W})} \end{bmatrix},$$

where  $J_{\psi}^{(q)}$  has  $d\binom{R+d_0-1}{R}$  columns and  $J_{\psi}^{(W)}$  has nd columns.

Now assume that there exists  $m \leq d_0$  and two matrices  $\mathbf{V} \in \mathbb{R}^{d \times m}$  and  $\mathbf{W} \in \mathbb{R}^{n \times d}$  such that the equalities (9)–(10) are satisfied (for example, if these are generic matrices from Lemma B.1); also, consider the vector polynomial in  $\mathbf{q}_0(x_1, \ldots, x_m) \in (\mathscr{H}_{m,R})^d \subseteq (\mathscr{H}_{d_0,R})^d$  defined as

$$\boldsymbol{q}_{0}(\boldsymbol{x}) := \boldsymbol{V} \begin{bmatrix} x_{1}^{R} \\ x_{2}^{R} \\ \vdots \\ x_{m}^{R} \end{bmatrix}. \tag{14}$$

Then we have that the evaluation of the Jacobian at the particular point  $(q_0, W)$  is of maximal possible rank, and, in particular,

$$\operatorname{rank}\{J_{\psi}(\boldsymbol{q}_{0},\boldsymbol{W})\} = d(n-1) + d\binom{R+d_{0}-1}{R}$$
(15)

and

$$\operatorname{rank}\{J_{\psi}^{(q)}(q_0, \boldsymbol{W})\} = d\binom{R + d_0 - 1}{R}$$
(16)

(i.e. the first block is full column rank).

Before proving Proposition B.5, we give an illustrative example of the proposition specializing to  $m = d_0 = 2, n = 1.$ 

**Example B.6.** In the notation of Example B.3, the vector polynomial  $q_0$  reads

$$\mathbf{q}_{0}(x_{1}, x_{2}) = \begin{bmatrix} (\alpha_{1}x_{1}^{R} + \beta_{1}x_{2}^{R})^{r} \\ (\alpha_{2}x_{1}^{R} + \beta_{2}x_{2}^{R})^{r} \\ \vdots \\ (\alpha_{d}x_{1}^{R} + \beta_{d}x_{2}^{R})^{r} \end{bmatrix},$$

and therefore the columns of  $J_{\psi}^{(q)}(q_0, \mathbf{W})$  (up to scaling and permutation) are the following polynomials:

$$f_{i,\ell}(x_1, x_2) := (\alpha_j x_1^R + \beta_j x_2^R)^{r-1} x_1^{R-\ell} x_2^{\ell}, \tag{17}$$

where  $j=1,\ldots,d$  and  $\ell=0,\ldots,R$ . So the proposition in this particular case proves that this set is linearly independent.

## **B.3** Proof of the key proposition: reducing the number of variables

We formulate the following lemma that tells us that we can always consider the case  $d_0 = m$  in the proof of Proposition B.5 or similar propositions.

**Lemma B.7.** Let  $d_0, d, n$  and  $r, R \geq 2$ , W be as in the statement of Proposition B.5, and for  $2 \leq m \leq d_0$  define the following submatrix of  $J_{3b}^{(q)}$ ,

$$J_{\psi}^{(q|_{1:m})}$$

which contains derivatives only with respect to the coefficients of  $\mathbf{q}$  for monomials  $x_1^{i_1}\dots x_m^{i_m}$ ; the matrix  $J_{\psi}^{(\mathbf{q}|_{1:m})}$  has  $d\binom{m+R-1}{R}$  columns and has the same number of rows as  $J_{\psi}^{(\mathbf{q})}$ .

Let  $q_0(x_1,...,x_m) \in (\mathcal{H}_{m,R})^d \subseteq (\mathcal{H}_{d_0,R})^d$  be some polynomial depending only on first m variables (not necessarily of the form (14)). and assume that the ranks of the reduced Jacobians satisfy:

$$rank\{\left[J_{\psi}^{(q|_{1:m})}(q_0, \mathbf{W}) \quad J_{\psi}^{(\mathbf{W})}\right]\} = d(n-1) + d\binom{R+m-1}{R}.$$
 (18)

and  $J_{\psi}^{(m{q}|_{1:m})}(m{q}_0, m{W})$  is full column rank:

$$\operatorname{rank}\{J_{\psi}^{(\boldsymbol{q}|_{1:m})}(\boldsymbol{q}_{0},\boldsymbol{W})\} = d \binom{R+m-1}{R}. \tag{19}$$

Then the full Jacobians  $J_{\psi}$  and  $J_{\psi}^{(q)}$  satisfy (15)–(16) (for a larger number of variables).

*Proof.* We first let  $W = [w_1 \quad \cdots \quad w_d]$  as in Remark B.4, so we can express

$$\psi[\boldsymbol{q}, \boldsymbol{W}] = \sum_{j=1}^{d} \boldsymbol{w}_{j}(q_{j})^{r}.$$

Already this, similarly to (12) gives us

$$\frac{\partial}{\partial W_{i,j}} \psi = \frac{\partial}{\partial (\boldsymbol{w}_j)_i} \psi = \boldsymbol{e}_i(q_j)^r,$$

we denote the space of these polynomials as  $\mathcal{L}^{(W)}$ .

We first look into details of the structure of the matrix  $J_{\psi}^{(q)}$ . Let  $i = (i_1, \dots, i_{d_0}) \in \mathcal{I}$  be a multi-index that runs over

$$\mathcal{I} = \{ \boldsymbol{i} := (i_1, \dots, i_{d_0}) : i_1, \dots, i_{d_0} \ge 0 \text{ and } i_1 + \dots + i_{d_0} = R \}$$

so that the coefficients of a polynomial  $q \in \mathscr{H}_{d_0,r}$  can be numbered by the elements in  $\mathcal{I}$ 

$$q(x_1,\ldots,x_{d_0}) = \sum_{i\in\mathcal{I}} q^{(i)} x_1^{i_1} \ldots x_{d_0}^{i_{d_0}}.$$

Then the columns of  $J_{\psi}^{(q)}$  for  $q(x) = [q_1(x) \cdots q_d(x)]^{\mathsf{T}}$  are given by the polynomials

$$\boldsymbol{f}_{j,\boldsymbol{i}}(\boldsymbol{x}) := \frac{\partial}{\partial q_j^{(\boldsymbol{i})}} \psi(\boldsymbol{q}, \boldsymbol{W}) = (rx_1^{i_1} \cdots x_{d_0}^{i_{d_0}}) \boldsymbol{w}_j(q_{0,j})^{r-1}, \quad j = 1, \dots, d, \quad \boldsymbol{i} \in \mathcal{I}, \quad (20)$$

where the last equality is similar to the one in (13). We denote spaces spanned by of such polynomials as

$$\mathcal{L}^{(q,i)} := \operatorname{span}\{f_{i,i}(x)\}_{i=1}^d$$
.

Now, consider a particular choice of  $q = q_0(x_1, \dots, x_m)$  depending only on the m variables. Then thanks to (20) we have

$$f_{j,(i_1,\ldots,i_{d_0})}(x_1,\ldots,x_{d_0}) = \underbrace{(\ldots)}_{\text{polynomial in } x_1,\ldots,x_m} x_{m+1}^{i_{m+1}}\ldots x_{d_0}^{i_{d_0}}.$$

Therefore, we get that  $\mathcal{L}^{(q,i)} \perp \mathcal{L}^{(q,\ell)}$  if  $(i_{m+1},\ldots,i_{d_0}) \neq (\ell_{m+1},\ldots,\ell_{d_0})$ 

Note that the extra columns (contained in  $J_{\psi}^{(q)}$  but not in  $J_{\psi}^{(q|_{1:m})}$  ) span the following subspace:

$$\mathcal{L}_{ext} := \bigoplus_{\substack{\boldsymbol{i} \in \mathcal{I}, \\ i_{m+1} + \dots + i_{d_0} > 0}} \mathcal{L}^{(\boldsymbol{q}, \boldsymbol{i})}.$$

To show that this space has maximal dimension, we consider splitting subsets  $\overline{\mathcal{I}_k}$ ,  $1 < k \le R$ :

$$\overline{\mathcal{I}}_k := \{(i_{m+1}, \dots, i_{d_0}) : i_k \ge 0, i_{m+1} + \dots + i_{d_0} = k\},\$$

Note that by orthogonality

$$\dim \mathcal{L}_{ext} = \sum_{k=1}^{R} \sum_{\substack{(i_{m+1}, \dots, i_{d_0}) \in \overline{\mathcal{I}}_k}} \dim \underbrace{\left( \bigoplus_{\substack{(i_1, \dots, i_m) : i_k \ge 0, \\ i_1 + \dots + i_m = R - k}} \mathcal{L}^{(q, (i_1, \dots, i_m))} \right)}_{=: \mathcal{M}_{(i_1, \dots, i_m)}}.$$

But then the dimension of the subspace  $\mathcal{M}_{(i_1,\ldots,i_m)}$  is equal to

$$\dim \mathcal{M}_{(i_1,\dots,i_m)} = \dim \operatorname{span}\{x_1^{k+i_1} \cdots x_m^{i_m} \boldsymbol{w}_j(q_{0,j})^{r-1}\}_{\substack{j=1,\dots,d,(i_1,\dots,i_m):\\i_k \geq 0, k+i_1+\dots+i_m=R}},$$

but the latter set of polynomials is linearly independent because it is spanned a subset of columns of  $J_{\psi}^{(q|_{1:m})}$ , which itself is full column rank by assumption (from (19)). Therefore we get  $\mathcal{L}_{ext}$  is of maximal possible dimension (the spanning columns are linearly independent).

But now recall that  $\mathcal{L}_{ext}$  is orthogonal both to span  $J_{\psi}^{(q|_{1:m})}$  and span  $J_{\psi}^{(W)}$ , hence (19) and (18) imply (16) and (15).

# **B.4** Proof of the key proposition: proof for $m=d_0$

*Proof of Proposition B.5.* Thanks to Lemma B.7, we can only consider the case  $m=d_0$  and prove (19) and (18) instead. Recall that in the notation of Lemma B.7, we need to calculate

$$\dim(\mathcal{L}^{(\boldsymbol{W})} \oplus \operatorname{span}\{\mathcal{L}^{(\boldsymbol{q},\boldsymbol{i})}\}_{\boldsymbol{i}\in\mathcal{I}})$$

Now let us consider these subspaces for a particular choice of  $q = q_0$  of the form (14).

$$f_{j,(i_1,\ldots,i_m)}(x_1,\ldots,x_m) = \underbrace{(\ \ \cdots\ \ )}_{\text{polynomial in } x_1^R,\ldots,x_m^R} x_1^{i_1(\bmod R)}\ldots x_m^{i_m(\bmod R)}.$$

Therefore we get that  $\mathcal{L}^{(q,i)} \perp \mathcal{L}^{(q,\ell)}$  unless one of the following conditions holds:

$$i = \ell$$
 or  $\{i, \ell\} \subset \mathcal{I}_0$ 

with  $\mathcal{I}_0 := \{(R,0,\ldots,0), (0,R,0,\ldots,0),\ldots,(0,0,\ldots,R)\}$ . For the same reasons we get  $\mathcal{L}^{(\boldsymbol{W})} \perp \mathcal{L}^{(\boldsymbol{q},\boldsymbol{i})} \text{ for all } \boldsymbol{i} \in \mathcal{I} \setminus \mathcal{I}_0.$ 

Therefore, we get

$$\operatorname{rank}\{J\} = \dim \left(\mathcal{L}^{(\boldsymbol{W})} \oplus \operatorname{span}\{\mathcal{L}^{(\boldsymbol{q},\boldsymbol{i})}\}_{\boldsymbol{i} \in \mathcal{I}_0}\right) + \sum_{\boldsymbol{i} \in \mathcal{I} \setminus \mathcal{I}_0} \dim(\mathcal{L}^{(\boldsymbol{q},\boldsymbol{i})}).$$

Let's look at those dimensions separately. Denote  $z = \begin{bmatrix} z_1 & \cdots & z_m \end{bmatrix}^T$ 

$$z_1 = x_1^R, \quad \dots, \quad z_m = x_m^R.$$

so that

$$q_{0,j} = \boldsymbol{v}_{i}^{\mathsf{T}} \boldsymbol{z}.$$

Then, for  $i \in \mathcal{I} \setminus \mathcal{I}_0$  it is easy to see that

$$\dim(\mathcal{L}^{(q,i)}) = \dim(\operatorname{span}\{\{\boldsymbol{w}_j(q_{0,j})^{r-1}\}_{j=1}^d\}) = d,$$

where the last equality follows from Lemma B.1 and (13).

By doing the same substitution, we obtain that

$$\mathcal{L}^{(W)} \oplus \text{span}\{\mathcal{L}^{(q,i)}\}_{i \in \mathcal{I}_0} = \text{span}\{\{e_i(v_j^{\mathsf{T}} z)^r\}_{i,j}^{n,d}\}, \{w_j z_\ell(v_j^{\mathsf{T}} z)^{r-1}\}_{j,\ell=1}^{d,m}\}\},$$

which is exactly the set of vectors in (13)–(12). Therefore, by Lemma B.1, we have

$$\dim \operatorname{span}\{\mathcal{L}^{(W)}, \{\mathcal{L}^{(q,\ell)}\}_{\ell \in \mathcal{I}_0}\} = (n-1)d + md, \text{ and}$$
(21)

$$\dim \operatorname{span}\{\mathcal{L}^{(q,\ell)}\}_{\ell\in\mathcal{I}_0} = md. \tag{22}$$

Taking into account that

$$\#(\mathcal{I}_0) = m$$
 and  $\#(\mathcal{I}) = {R+m-1 \choose R}$ ,

this proves (15) for  $d_0 = m$ . Equality (16) (also for  $d_0 = m$ ) can be proved similarly using the fact that

$$\begin{split} \operatorname{rank} \{J_{\psi}^{(\boldsymbol{q})}(\boldsymbol{q}_0, \boldsymbol{W})\} &= \operatorname{dim} \operatorname{span} \{\mathcal{L}^{(\boldsymbol{q}, \boldsymbol{\ell})}\}_{\boldsymbol{\ell} \in \mathcal{I}_0} + \sum_{\boldsymbol{i} \in \mathcal{I} \backslash \mathcal{I}_0} \operatorname{dim} (\mathcal{L}^{(\boldsymbol{q}, \boldsymbol{i})}) \\ &= md + d(\#(\mathcal{I}) - \#(\mathcal{I}_0)) = d(\#(\mathcal{I})). \end{split}$$

# **B.5** An illustration: the bivariate case

To fix the ideas, we consider the case  $d_0=2$  and n=1,  $\boldsymbol{W}=\begin{bmatrix}1&\cdots&1\end{bmatrix}$  so we are in the notation of Example B.3. In this case, as before, we would be looking only at  $J_{\psi}^{(\boldsymbol{q})}(\boldsymbol{q}_0,\boldsymbol{W})$  (since  $\mathcal{L}^{(\boldsymbol{W})}$  does not add new information. In this case, as in Example B.6, we look at the span of  $f_{j,\ell}$  defined in (17) for  $j=1,\ldots,d$  and  $\ell\in\{0,\ldots,R\}$ .

With abuse of notation, we will split similarly into subspaces as  $\mathcal{L}^{(q)} = \mathcal{L}^{(q,0)} \oplus \cdots \oplus \mathcal{L}^{(q,R)}$ 

$$\mathcal{L}^{(q,\ell)} := \operatorname{span}\{f_{j,\ell}\}_{j=1}^d.$$

Note that  $\dim \mathcal{L}^{(q,\ell)} = \dim \mathcal{L}^{(q,i)}$  for all  $i, \ell$ . But then we have that

$$f_{j,\ell}(x_1,x_2) := \underbrace{(\cdots)}_{\text{polynomial in } x_1^R,x_2^R} x_1^{(R-\ell)(\bmod{R})} x_2^{\ell(\bmod{R})}$$

Therefore  $\mathcal{L}^{(q,i)} \perp \mathcal{L}^{(q,\ell)}$  unless  $i \pmod{R} = \ell \pmod{R}$ , which implies

$$\dim(\mathcal{L}^{(q)}) = \dim(\mathcal{L}^{(q,0)} \oplus \mathcal{L}^{(q,R)}) + (R-1) \cdot \dim(\mathcal{L}^{(q,0)}).$$

To get the dimension of span $\{\mathcal{L}^{(q,0)},\mathcal{L}^{(q,R)}\}\$ , observe that this space is spanned by 2d polynomials

$$(\alpha_j x_1^R + \beta_j x_2^R)^{r-1} x_1^R, (\alpha_j x_1^R + \beta_j x_2^R)^{r-1} x_2^R, \quad j = 1, \dots, d.$$

where only multiples of R appear as powers  $z_1 = x_1^R$  and  $z_2 = x_2^R$ , then this will be exactly the set of polynomials

$$z_1(\alpha_j z_1 + \beta_j z_2)^{r-1}, z_2(\alpha_j z_1 + \beta_j z_2)^{r-1},$$

which are shown in Example B.3 to be linearly independent if  $r \geq 2d-1$  and  $(\alpha_j, \beta_j)$  non-collinear. This proves that  $\dim(\operatorname{span}\{\mathcal{L}^{(q,0)}, \mathcal{L}^{(q,R)}\}) = 2d$ , which would imply  $\dim(\mathcal{L}^{(q,0)}) = d$ . Combining it all together, we get

$$\dim(\mathcal{L}^{(q)}) = 2d + (R-1)d = (R+1)d.$$

# **B.6** Localization theorem

**Theorem 11 (Localization theorem)** Let  $((d_0,\ldots,d_L),(r_1,\ldots,r_{L-1}))$  be the hPNN format. For  $\ell=0,\ldots,L-2$  denote  $\widetilde{d}_\ell=\min\{d_0,\ldots,d_\ell\}$ . Then the following holds true: if for all  $\ell=1,\ldots,L-1$  the two-layer architecture  $\operatorname{hPNN}_{(\widetilde{d}_{\ell-1},d_\ell,d_{\ell+1}),r_\ell}[\cdot]$  is finitely identifiable, then the L-layer architecture  $\operatorname{hPNN}_{d,r}[\cdot]$  is finitely identifiable as well.

*Proof.* (Proof of Theorem 11) We prove the theorem by induction.

- Base: L=2 The base of the induction is trivial since the case L=2 the full hPNN consists in a 2-layer network.
- Induction step:  $(L = k 1) \rightarrow (L = k)$  Assume that the statement holds for L = k 1. Now consider the case L = k.

Let 
$$\boldsymbol{\theta} = (\boldsymbol{W}_1, \dots, \boldsymbol{W}_{L-1})$$
, so that  $\boldsymbol{w} = (\boldsymbol{\theta}, \boldsymbol{W}_L)$  and denote  $R = r_1 \cdots r_{L-2}$ .

Let  $\psi$  be as the one defined in Proposition B.5, but given for the last subnetwork, so that  $n=d_L, d=d_{L-1}, r=r_{L-1}, \boldsymbol{W}=\boldsymbol{W}_L$ . Then we have that

$$\boldsymbol{p}_{\boldsymbol{w}} = \text{hPNN}_{(r_1, \dots, r_{L-1})}[(\boldsymbol{\theta}, \boldsymbol{W}_L)] = \psi[\phi(\boldsymbol{\theta}), \boldsymbol{W}_L]$$

where 
$$\phi(\boldsymbol{\theta}) = \text{hPNN}_{(r_1, \dots, r_{L-2})}[\boldsymbol{\theta}].$$

Therefore, by the chain rule

$$egin{aligned} oldsymbol{J_{p_w}}(w) = \left[ \underbrace{egin{aligned} oldsymbol{J_{\psi}}^{(q)} \Big|_{q=\phi( heta)} & oldsymbol{J_{\psi}}^{(W)} \Big|_{q=\phi( heta)} \ & \underbrace{oldsymbol{J_{\psi}}^{(W)} \Big|_{q=\phi( heta)}}_{=oldsymbol{J_{2}( heta)}} 
ight], \end{aligned}$$

Now we are going to show that the matrices have necessary rank for generic  $\theta$ . For this, note by the induction assumption, for generic  $\theta$ , we have

$$\operatorname{rank}\{\boldsymbol{J}_{\phi}(\boldsymbol{\theta})\} = \sum_{\ell=0}^{L-2} d_{\ell} d_{\ell+1} - \sum_{\ell=1}^{L-2} d_{\ell}.$$

Now we show the ranks for other matrices. Observe that

$$\operatorname{rank}\left\{\left[\left(\boldsymbol{J}_{\psi}^{(q)}\Big|_{\boldsymbol{q}=\phi(\boldsymbol{\theta})}\right) \quad \boldsymbol{J}_{\psi}^{(\boldsymbol{W})}\Big|_{\boldsymbol{q}=\phi(\boldsymbol{\theta})}\right]\right\} \leq d_{L-1}\binom{R+d_0-1}{R} + (d_L-1)d_{L-1} \quad (23)$$

due to the essential ambiguities. But then if we find a particular point  $\hat{\theta}$ , where rank is maximal for  $q_0 = \phi(\hat{\theta})$ , then the rank in (23) will be maximal for generic  $\theta$ .

But then, let  $m = \tilde{d}_{L-1} = \min\{d_0, \dots, d_{L-1}\}$  and consider the following matrices:

$$egin{aligned} \widehat{m{W}}_1 = egin{bmatrix} m{I}_m & \mathbf{0} \ \mathbf{0} & \mathbf{0} \end{bmatrix}, \dots, \widehat{m{W}}_{L-2} = egin{bmatrix} m{I}_m & \mathbf{0} \ \mathbf{0} & \mathbf{0} \end{bmatrix} \end{aligned}$$

and

$$\widehat{\boldsymbol{W}}_{L-1} = [\boldsymbol{V} \quad \boldsymbol{0}],$$

for  $m{V} \in \mathbb{R}^{d_{L-1} imes m}$  generic. Then we get that for  $\widehat{m{ heta}} = (\widehat{m{W}}_1, \dots, \widehat{m{W}}_{L-2})$ 

$$\phi(\widehat{\boldsymbol{\theta}}) = \boldsymbol{V} \begin{bmatrix} x_1^R \\ \vdots \\ x_m^R \end{bmatrix},$$

so exactly as in Proposition B.5. Therefore rank in (23) will be maximal for generic  $(\theta, W_L)$  and also

$$\operatorname{rank}\{\left(J_{\psi}^{(\boldsymbol{q})}\Big|_{\boldsymbol{q}=\phi(\boldsymbol{\theta})}\right)\}=d_{L-1}\binom{R+d_0-1}{R}$$

for generic  $\theta$  (i.e., the matrix is full rank).

This leads to rank $\{J_1(w)\} = J_{\phi}(\theta)$  for generic  $\theta$ . Finally, we have that

$$\begin{aligned} \operatorname{rank}\{\boldsymbol{J}_{\boldsymbol{p_{w}}}(\boldsymbol{w})\} &= \operatorname{rank}\{\boldsymbol{J}_{1}(\boldsymbol{\theta})\} + \operatorname{rank}\{\boldsymbol{\Pi}_{\operatorname{span}}\boldsymbol{J}_{1}(\boldsymbol{\theta})_{\perp}\boldsymbol{J}_{2}(\boldsymbol{\theta})\} \\ &\geq \operatorname{rank}\{\boldsymbol{J}_{1}(\boldsymbol{\theta})\} + \operatorname{rank}\{\boldsymbol{\Pi}_{\operatorname{span}}\left(\boldsymbol{J_{\psi}^{(\boldsymbol{q})}}\Big|_{\boldsymbol{q}=\phi(\boldsymbol{\theta})}\right)_{\perp}\boldsymbol{J}_{2}(\boldsymbol{\theta})\} \\ &= \sum_{\ell=0}^{L-2} d_{\ell}d_{\ell+1} - \sum_{\ell=1}^{L-2} d_{\ell} + (d_{L}-1)d_{L-1} \\ &= \sum_{\ell=0}^{L-1} d_{\ell}d_{\ell+1} - \sum_{\ell=1}^{L-1} d_{\ell}\,, \end{aligned}$$

where  $\Pi_{\mathcal{U}}$  denotes the orthogonal projection onto some subspace  $\mathcal{U}$ . On the other hand,

$$\operatorname{rank}\{\boldsymbol{J}_{\boldsymbol{p}_{\boldsymbol{w}}}(\boldsymbol{w})\} \leq \sum_{\ell=0}^{L-1} d_{\ell} d_{\ell+1} - \sum_{\ell=1}^{L-1} d_{\ell}$$

due to presence of ambiguities. Hence, an equality holds and therefore the neurovariety has expected dimension.

# B.7 Implications of the localization theorem

**Corollary 16 (Pyramidal)** The hPNNs with architectures containing non-increasing layer widths  $d_0 \geq d_1 \geq \cdots d_{L-1} \geq 2$ , except possibly for  $d_L \geq 1$  are finitely identifiable for any degrees satisfying (i)  $r_1, \ldots, r_{L-1} \geq 2$  if  $d_L \geq 2$ ; or (ii)  $r_1, \ldots, r_{L-2} \geq 2$ ,  $r_{L-1} \geq 3$  if  $d_L \geq 1$ .

*Proof.* (Proof of Corollary 16) This follows from the following facts:

- For such a choice of  $d_{\ell}$ ,  $\widetilde{d}_{\ell} = d_{\ell}$  for all  $\ell = 0, \dots, L-1$ ;
- Network  $(d_{\ell-1}, d_{\ell}, d_{\ell+1})$  with  $d_{\ell-1} \ge d_{\ell}$  is identifiable for:
  - $r_{\ell} \ge 2$ , in case  $d_{\ell+1} \ge 2$ ;
  - $r_{\ell} \ge 3$ , in case  $d_{\ell+1} = 1$ .

Corollary 17 (Activation thresholds for identifiability) For fixed layer widths  $\mathbf{d} = (d_0, \dots, d_L)$  with  $d_\ell \geq 2, \ell = 0, \dots, L-1$ , the hPNNs with architectures  $(\mathbf{d}, (r_1, \dots, r_{L-1}))$  are identifiable for any degrees satisfying

$$r_{\ell} \geq 2d_{\ell} - 1$$
.

*Proof.* (proof of Corollary 17) We observe that this guarantees that  $\widetilde{d}_{\ell} \geq 2$ . But then the Kruskal bound for identifiability of  $(\widetilde{d}_{\ell-1}, d_{\ell}, d_{\ell+1})$  is

$$\frac{2d_{\ell} - \min(d_{\ell}, d_{\ell+1})}{\min(d_{\ell}, \widetilde{d}_{\ell-1}) - 1} \le 2d_{\ell} - 1.$$

therefore, for  $r_{\ell} \geq 2d_{\ell} - 1$  the hPNN  $(\widetilde{d}_{\ell-1}, d_{\ell}, d_{\ell+1}), r_{\ell}$  is identifiable.

Corollary 19 (Identifiability of bottleneck hPNNs) Consider the "bottleneck" architecture with

$$d_0 \geq d_1 \geq \cdots \geq d_b \leq d_{b+1} \leq \ldots \leq d_L$$

and  $d_b \geq 2$ . Suppose that  $r_1, \ldots, r_b \geq 2$  and that the decoder part satisfies  $\frac{d_\ell}{r_\ell} \leq d_b - 1$  for  $\ell \in \{b+1, \ldots, L-1\}$ . Then the bottleneck hPNN is finitely identifiable.

Proof. (proof of Corollary 19) This follows from Theorem 11 and the following facts:

- For layers  $\ell \in \{1, \dots, b\}$  (the encoder part), we have  $\widetilde{d}_{\ell} = d_{\ell}$  and thus identifiability of  $(\widetilde{d}_{\ell-1}, d_{\ell}, d_{\ell+1})$  holds for  $r_{\ell} \geq 2$  (the same argument as in the pyramidal case).
- For layers  $\ell \in \{b+1,\ldots,L\}$  (the decoder part), we have  $\widetilde{d}_{\ell}=d_b$  and thus identifiability of  $(\widetilde{d}_{\ell-1},d_{\ell},d_{\ell+1})$  holds for

$$r_{\ell} \geq \frac{d_{\ell}}{d_{h}-1}$$

rearranging gives the desired result.

# C Analyzing case of PNNs with biases

This appendix contains the proofs and supporting technical results for the identifiability results of PNNs with bias terms presented in Section 3.3 of the main paper. We start by establishing the relationship between PNNs and hPNNs and their uniqueness by means of homogeneization. We then prove our main finite identifiability results showing that finite identifiability of 2-layer subnetworks of the homogeneized PNNs is sufficient to guarantee the finite identifiability of the original PNN.

**Results from the main paper**: Definition 20, Propositions 23, 24, and 27, Lemma 26, and Corollary 28.

# C.1 The homogeneization procedure: the hPNN associated to a PNN

Our homogeneization procedure is based on the following lemma:

**Definition 20.** There is a one-to-one mapping between (possibly inhomogeneous) polynomials in d variables of degree r and homogeneous polynomials of the same degree in d+1 variables. We denote this mapping  $\mathscr{P}_{d,r} \to \mathscr{H}_{d+1,r}$  by  $homog(\cdot)$ , and it acts as follows: for every polynomial  $p \in \mathscr{P}_{d,r}$ ,  $\widetilde{p} = homog(p) \in \mathscr{H}_{d+1,r}$  is the unique homogeneous polynomial in d+1 variables such that

$$\widetilde{p}(x_1,\ldots,x_d,1)=p(x_1,\ldots,x_d).$$

Proof of Definition 20. Let p be a possibly inhomogeneous polynomial in d variables, which reads

$$p(x_1, \dots, x_d) = \sum_{\alpha, |\alpha| \le r} b_{\alpha} x_1^{\alpha_1} \cdots x_d^{\alpha_d},$$

for  $\alpha = (\alpha_1, \dots, \alpha_d)$ . One sets

$$\widetilde{p}(x_1, \dots, x_d, z) = \sum_{\alpha, |\alpha| \le r} b_{\alpha} x_1^{\alpha_1} \cdots x_d^{\alpha_d} z^{r - \alpha_1 - \dots - \alpha_d}$$

which satisfies the required properties.

Associating an hPNN to a given PNN: Now we prove that for each polynomial p admitting a PNN representation, its associated homogeneous polynomial admits an hPNN representation. This is formalized in the following result.

**Proposition 23.** Fix the architecture  $\mathbf{r} = (r_1, \dots, r_L)$  and  $\mathbf{d} = (d_0, \dots, d_L)$ . Then a polynomial vector  $\mathbf{p} \in (\mathscr{P}_{d_0, r_{total}})^{\times d_L}$  admits a PNN representation  $\mathbf{p} = \text{PNN}_{\mathbf{d}, \mathbf{r}}[(\mathbf{w}, \mathbf{b})]$  with  $(\mathbf{w}, \mathbf{b})$  as in (2) if and only if its homogeneization  $\widetilde{\mathbf{p}} = \text{homog}(\mathbf{p})$  admits an hPNN decomposition for the same activation degrees  $\mathbf{r}$  and extended  $\widetilde{\mathbf{d}} = (d_0 + 1, \dots, d_{L-1} + 1, d_L)$ ,  $\widetilde{\mathbf{p}} = \text{hPNN}_{\widetilde{\mathbf{d}}, \mathbf{r}}[\widetilde{\mathbf{w}}]$ ,  $\widetilde{\mathbf{w}} = (\widetilde{\mathbf{W}}_1, \dots, \widetilde{\mathbf{W}}_{L-1}, [\mathbf{W}_L \quad \mathbf{b}_L])$ , with matrices  $\widetilde{\mathbf{W}}_\ell$  for  $\ell = 1, \dots, L-1$  given as

$$\widetilde{\boldsymbol{W}}_{\ell} = \begin{bmatrix} \boldsymbol{W}_{\ell} & \boldsymbol{b}_{\ell} \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{(d_{\ell}+1) \times (d_{\ell-1}+1)}$$
.

Proof of Proposition 23. Denote  $p_1(x) = \rho_{r_1}(W_1x + b_1)$ . Let  $\widetilde{x} = \begin{bmatrix} x \\ z \end{bmatrix} \in \mathbb{R}^{d_0+1}$ . Observe first that

$$ho_{r_1}(\widetilde{oldsymbol{W}}_1\widetilde{oldsymbol{x}}) = \left[egin{matrix} \widetilde{oldsymbol{p}}_1(\widetilde{oldsymbol{x}}) \ z^{r_1} \end{array}
ight].$$

We proceed then by induction on  $L \geq 1$ .

The case L=1 is trivial. Assume that L=2. Then

$$\widetilde{\boldsymbol{W}}_2 \rho_{r_1}(\widetilde{\boldsymbol{W}}_1 \widetilde{\boldsymbol{x}}) = \widetilde{\boldsymbol{W}}_2 \begin{bmatrix} \widetilde{\boldsymbol{p}}_1(\widetilde{\boldsymbol{x}}) \\ z^{r_1} \end{bmatrix} = \boldsymbol{W}_2 \widetilde{\boldsymbol{p}}_1(\widetilde{\boldsymbol{x}}) + z^{r_1} \boldsymbol{b}_2.$$

Specializing at z = 1, we recover

$$W_2 p_1(x) + b_2 = p(x) = \widetilde{p}(x, 1)$$

hence

$$\widetilde{\boldsymbol{W}}_2 \rho_{r_1}(\widetilde{\boldsymbol{W}}_1 \widetilde{\boldsymbol{x}}) = \widetilde{\boldsymbol{p}}(\widetilde{\boldsymbol{x}}).$$

For the induction step, assume that  $\tilde{q} = \text{hPNN}_{(d_1+1,\dots,d_{L-1}+1,d_L),r}[(\widetilde{\boldsymbol{W}}_2,\dots,\widetilde{\boldsymbol{W}}_L)]$  is the homogeneization of  $q = \text{PNN}_{(d_1,\dots,d_L),r}[((\boldsymbol{W}_2,\dots,\boldsymbol{W}_L),(\boldsymbol{b}_2,\dots,\boldsymbol{b}_L))]$ . By assumption,

$$\widetilde{\boldsymbol{p}}(\boldsymbol{x},1) = \widetilde{\boldsymbol{q}}\left(\begin{bmatrix}\widetilde{\boldsymbol{p}}_1(\boldsymbol{x},1)\\1\end{bmatrix}\right) = \boldsymbol{q}(\widetilde{\boldsymbol{p}}_1(\boldsymbol{x},1)) = \boldsymbol{q}(\boldsymbol{p}_1(\boldsymbol{x})) = \boldsymbol{p}(\boldsymbol{x})\,.$$

**Proposition 24.** If  $hPNN_r[\widetilde{w}]$  from Proposition 23 is unique as an hPNN (without taking into account the structure), then the original PNN representation  $PNN_r[(w, b)]$  is unique.

Proof of Proposition 24. Suppose  $\operatorname{hPNN}_r[\widetilde{w}]$  is unique (or finite-to-one), where  $\widetilde{w}$  is structured as in Proposition 23. Note that any equivalent (in the sense of Lemma 4 specialized for  $\operatorname{hPNN}_r[\widetilde{w}]$ ) parameter vector  $\widetilde{w}' = (\widetilde{\boldsymbol{W}}_1', \dots, \widetilde{\boldsymbol{W}}_L')$  realizing the same hPNN must satisfy

$$\widetilde{\boldsymbol{W}}_{\ell}' = \begin{cases} \widetilde{\boldsymbol{P}}_{\ell} \widetilde{\boldsymbol{D}}_{\ell} \widetilde{\boldsymbol{W}}_{\ell} \widetilde{\boldsymbol{D}}_{\ell-1}^{-r_{\ell-1}} \widetilde{\boldsymbol{P}}_{\ell-1}^{\mathsf{T}}, & \ell < L, \\ \widetilde{\boldsymbol{W}}_{L} \widetilde{\boldsymbol{D}}_{L-1}^{-r_{L-1}} \widetilde{\boldsymbol{P}}_{L-1}^{\mathsf{T}}, & \ell = L. \end{cases}$$

$$(24)$$

for permutation matrices  $\widetilde{\boldsymbol{P}}_{\ell}$  and invertible diagonal matrices  $\widetilde{\boldsymbol{D}}_{\ell}$ , with  $\widetilde{\boldsymbol{P}}_0 = \widetilde{\boldsymbol{D}}_0 = \boldsymbol{I}$ . We are going to show that bringing  $\widetilde{\boldsymbol{W}}'_{\ell}$  to the form

$$\widetilde{\boldsymbol{W}}_{\ell}' = \begin{cases} \begin{bmatrix} \boldsymbol{W}_{\ell}' & \boldsymbol{b}_{\ell}' \\ 0 & 1 \end{bmatrix}, & \ell < L, \\ \begin{bmatrix} \boldsymbol{W}_{L}' & \boldsymbol{b}_{L}' \end{bmatrix}, & \ell = L. \end{cases}$$
(25)

that does not introduce extra ambiguities besides the ones for PNN (given in Lemma 4).

Next, by Proposition 33, for  $\ell=1,\ldots,L-1$  the matrices satisfy  $\operatorname{krank}\{(\widetilde{\boldsymbol{W}}_{\ell})^{\mathsf{T}}\}\geq 2$  (as well as for any equivalent  $\operatorname{krank}\{(\widetilde{\boldsymbol{W}}'_{\ell})^{\mathsf{T}}\}\geq 2$ ). This implies that the matrix  $\widetilde{\boldsymbol{W}}_{\ell}$  contains only a single row of the form  $[0\cdots 0\,\alpha]$  (which is its last row). Therefore in order for  $\widetilde{\boldsymbol{W}}'_1$  to be of the form (25), the matrices  $\widetilde{\boldsymbol{P}}_1$   $\widetilde{\boldsymbol{D}}_1$  must be of the form

$$\widetilde{\boldsymbol{P}}_1 = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}, \qquad \widetilde{\boldsymbol{D}}_1 = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}.$$

Iterating this process for  $\ell = 2, \dots, L-1$ , we impose constraints of the form

$$\widetilde{\boldsymbol{P}}_{\ell} = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}, \qquad \widetilde{\boldsymbol{D}}_{\ell} = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}.$$

This implies that  $(\mathbf{W}'_{\ell}, \mathbf{b}'_{\ell})$  and  $(\mathbf{W}_{\ell}, \mathbf{b}_{\ell})$  must be linked as in Lemma 4.

Now suppose that  $\operatorname{hPNN}_r[\widetilde{w}]$  is finite-to-one. Then the same reasoning applies to all alternative (non-equivalent) parameters  $\widetilde{w}$  that are realized by a PNN, because Proposition 23 holds for every solution. Since there are finitely many equivalence classes, the corresponding PNN representation is also finite-to-one.

# C.2 Generic identifiability conditions for PNNs with bias terms

**Lemma 26** Let the 2-layer hPNN architecture  $((d_0 + 1, d_1 + 1, d_2), (r_1))$  be finitely (resp. globally) identifiable. Then the PNN architecture with widths  $(d_0, d_1, d_2)$  and degree  $r_1$  is also finitely (resp. globally) identifiable.

*Proof of Lemma 26.* By Proposition 24 we just need to show that for general  $(W_2, b_2, W_1, b_1)$ , the following hPNN is unique (finite-to-one)

$$p(\widetilde{x}) = [W_2 \quad b_2] \rho_{r_1}(\widetilde{W}_1 \widetilde{x})$$
(26)

with  $\widetilde{\boldsymbol{W}}_1$  given as

$$\widetilde{\boldsymbol{W}}_1 = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{b}_1 \\ 0 & 1 \end{bmatrix}.$$

We see that  $\widetilde{W}_1$  lies in a subspace of  $(d_1 + 1) \times (d_0 + 1)$  matrices.

We use the following fact: by multilinearity, both uniqueness and finite-to-one properties of an hPNN are invariant under multiplication of  $\widetilde{W}_1$  on the right by any nonsingular  $(d_0 + 1) \times (d_0 + 1)$  matrix Q. We note that the image of the polynomial map

$$\mathbb{R}^{(d_0+1)\times(d_0+1)}\times\mathbb{R}^{d_1\times d_0}\times\mathbb{R}^{d_0}\to\mathbb{R}^{(d_1+1)\times(d_0+1)}$$
$$(\boldsymbol{Q},\boldsymbol{W}_1,\boldsymbol{b}_1)\mapsto\widetilde{\boldsymbol{W}}_1\boldsymbol{Q},$$

which is surjective, and its image is dense.

Therefore, identifiability (resp. finite identifiability) holds except some set of measure zero in  $\mathbb{R}^{(d_1+1)\times(d_0+1)}$ , then it also hold for  $\widetilde{\boldsymbol{W}}_1$  constructed from almost all  $(\boldsymbol{W}_1,\boldsymbol{b}_1)$  pairs. For example, in terms of finite identifiability this is explained by the fact that there is a smooth point of the hPNN neurovariety corresponding to the parameters  $([\boldsymbol{W}_2 \quad \boldsymbol{b}_2], \widetilde{\boldsymbol{W}}_1)$ .

**Proposition 27** Let  $((d_0,\ldots,d_L),(r_1,\ldots,r_{L-1}))$  be the PNN format. For  $\ell=0,\ldots,L-2$  denote  $\widetilde{d}_\ell=\min\{d_0,\ldots,d_\ell\}$ . Then the following holds true: If for all  $\ell=1,\ldots,L-1$  the two layer architecture  $\operatorname{hPNN}_{(\widetilde{d}_{\ell-1}+1,d_{\ell+1}),r_\ell}[\cdot]$  is finitely identifiable, then the L-layer PNN with architecture  $(\boldsymbol{d},\boldsymbol{r})$  is finitely identifiable as well.

For the proof of the main proposition, we need the following lemma.

**Lemma C.1.** Global (resp. finite) identifiability of an hPNN of format ((m, d, n), r) implies (resp. finite) identifiability of the hPNN in format ((m, d, n + k), r) for any k > 0.

*Proof.* Let the parameters be such that

$$oldsymbol{W}_2 = egin{bmatrix} oldsymbol{A} \\ oldsymbol{B} \end{bmatrix}, \, oldsymbol{A} \in \mathbb{R}^{n imes d}, \, oldsymbol{B} \in \mathbb{R}^{k imes d}, \, oldsymbol{W}_1,$$

so that

$$\boldsymbol{p}_{(\boldsymbol{W}_1,\boldsymbol{W}_2)} = \begin{bmatrix} \boldsymbol{p}_{(\boldsymbol{W}_1,\boldsymbol{A})} \\ \boldsymbol{p}_{(\boldsymbol{W}_1,\boldsymbol{B})} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}\sigma_r(\boldsymbol{W}_1\boldsymbol{x}) \\ \boldsymbol{B}\sigma_r(\boldsymbol{W}_1\boldsymbol{x}) \end{bmatrix}.$$

But then assume that  $p^{(A)}$  is finite-to-one at  $(W_1, A)$  a given parameter. Then by Lemma 31 we have that the elements of  $q_1 = \sigma_r(W_1x)$  are linearly independent, hence B has a unique solution from the linear system

$$\boldsymbol{p}_{(\boldsymbol{W}_1,\boldsymbol{B})} = \boldsymbol{B}\sigma_r(\boldsymbol{W}_1\boldsymbol{x}).$$

Note that for  $(W_1, W_2)$ , the subset of parameters  $(W_1, A)$  is also generic, hence global (resp. finite) identifiability for widths (m, d, n) implies global (resp. finite) identifiability for widths (m, d, n + k).

*Proof of Proposition 27.* We are going to prove that under the condition of the theorem, two hPNN architectures for degrees r and widths

$$(d_0+1,\ldots,d_{L-1}+1,d_L)$$
 and  $(d_0+1,\ldots,d_{L-1}+1,d_L+1)$ 

are finitely identifiable.

We proceed by induction, similarly as in Theorem 11.

- Base: L=2 The base of the induction follows is trivial since it is the 2-layer network, and from Lemma C.1 for the architecture  $(d_{\ell-1}+1,d_{\ell}+1,d_{\ell+1}+1)$ .
- Induction step:  $(L=k-1) \to (L=k)$  Assume that the statement holds for L=k-1. Now consider the case L=k. As in the proof of Theorem 11, we set  $\widetilde{\boldsymbol{\theta}}=(\widetilde{\boldsymbol{W}}_1,\ldots,\widetilde{\boldsymbol{W}}_{L-1})$ , so that  $\widetilde{\boldsymbol{w}}=(\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{W}}_L)$  and denote  $R=r_1\cdots r_{L-2}$ . The difference is that the parameters are now  $\widetilde{\boldsymbol{\theta}}:=\widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$ , where

$$\theta = (W_1, \dots, W_{L-1}, b_1, \dots, b_{L-1}).$$

Let  $\psi$  be as the one defined in Proposition B.5, but given for the last subnetwork, so that  $n = d_L$ ,  $d = d_{L-1} + 1$ ,  $r = r_{L-1}$ ,  $W = \widetilde{W}_L$ . Then we have that

$$p_{\widetilde{\boldsymbol{w}}}(\widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta}), \boldsymbol{W}) = \psi[\phi(\widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta})), \boldsymbol{W}_L]$$

where  $\phi(\widetilde{\boldsymbol{\theta}}) = \text{hPNN}_{(r_1, \dots, r_{L-2})}[\widetilde{\boldsymbol{\theta}}].$ 

Again, by the chain rule

$$egin{aligned} oldsymbol{J_{p_{\widetilde{w}}}}(oldsymbol{w}) = \left[ \underbrace{egin{pmatrix} oldsymbol{J_{\psi}}^{(oldsymbol{q})} \Big|_{oldsymbol{q} = \phi(\widetilde{ heta}(oldsymbol{ heta}))} \end{pmatrix} \cdot oldsymbol{J_{\phi}}(\widetilde{ heta}(oldsymbol{ heta}))}{=oldsymbol{J_{1}}(\widetilde{oldsymbol{w}})} 
ight] egin{pmatrix} oldsymbol{J_{\psi}}(oldsymbol{ heta}) & oldsymbol{ heta}(oldsymbol{ heta}) & oldsymbol{J_{\psi}}(oldsymbol{ heta}) & oldsymbol{J_{\psi}}(oldsymbol{ heta}) & oldsymbol{ heta}(oldsymbol{ heta}) & oldsymbol{$$

where the matrix in the right hand side is full column rank. Therefore, we just need to show that the left hand side matrix is full column rank for a particular  $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$ . But for this remark that we can use almost the same construction example Proposition B.5, but choosing slightly different matrices:  $\widetilde{\boldsymbol{\theta}}' = (\widehat{\boldsymbol{W}}'_1, \dots, \widehat{\boldsymbol{W}}'_{L-1})$  with

$$\widehat{m{W}}_1' = egin{bmatrix} m{0} & m{0} \ m{0} & m{I}_m \end{bmatrix}, \dots, \widehat{m{W}}_{L-2} = egin{bmatrix} m{0} & m{0} \ m{0} & m{I}_m \end{bmatrix}$$

and

$$\widehat{\boldsymbol{W}}'_{L-1} = \begin{bmatrix} \mathbf{0} & \boldsymbol{V}' \end{bmatrix},$$

where in Lemma B.1 we can choose generic V' structured as

$$V' = \begin{bmatrix} W^{(V')} & b^{(V')} \\ 0 & 1 \end{bmatrix}.$$

Indeed, we need this to be a smooth point (i.e., full rank Jacobian of  $W\rho_{r_{L-1}}(V'x)$ ), which is full rank for generic  $W^{(V')}$ ,  $b^{(V')}$ , by the same argument as in the proof of Lemma 26.

But such  $\widetilde{\boldsymbol{\theta}}'$  indeed belongs to the image of  $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$  as they share the needed structure, which completes the proof.

**Corollary 28.** Let  $((d_0,\ldots,d_L),(r_1,\ldots,r_{L-1}))$  be such that  $d_\ell \geq 1$ , and  $r_\ell \geq 2$  satisfy

$$r_{\ell} \ge \frac{2(d_{\ell}+1) - \min(d_{\ell}+1, d_{\ell+1})}{\min(d_{\ell}, \tilde{d}_{\ell-1})},$$

then the L-layer PNN with architecture (d, r) is finitely identifiable (and globally identifiable when L = 2).

*Proof of Corollary 28.* This directly follows from combining Lemma 26, Proposition 12 and Proposition 27.  $\Box$ 

# D Homogeneous PNNs and neurovarieties

hPNNs are often studied through the prism of neurovarieties, using their algebraic structure. Our results have direct implications on the expected dimension of the neurovarieties. An hPNN hPNN<sub>r</sub>[w] can be equivalently defined by a map  $\Psi_{d,r}$  from the weight tuple w to a vector of homogeneous polynomials of degree  $r_{total} = r_1 r_2 \dots r_{L-1}$  in  $d_0$  variables:

$$\Psi_{\boldsymbol{d},\boldsymbol{r}}: \quad \boldsymbol{w} \mapsto \text{hPNN}_{\boldsymbol{d},\boldsymbol{r}}[\boldsymbol{w}] \\ \mathbb{R}^{\sum_{\ell} d_{\ell} d_{\ell-1}} \to (\mathscr{H}_{d_0,r_{total}})^{\times d_L}$$

where  $\mathscr{H}_{d,r} \subset \mathscr{P}_{d,r}$  denotes the space of homogeneous d-variate polynomials of degree r. The image of  $\Psi_{d,r}$  is called a *neuromanifold*, and the *neurovariety*  $\mathscr{V}_{d,r}$  is defined as its closure in Zariski topology (that is, the smallest algebraic variety that contains the image of the map  $\Psi_{d,r}$ ). Note that the neurovariety depends on the field (i.e., results can differ between  $\mathbb{R}$  or  $\mathbb{C}$ ), nonetheless, the following results hold for both the real and the complex case.

The key property of the neurovariety is its dimension, roughly defined as the dimension of the tangent space to a general point on the variety. The following upper bound was presented in [7]:

$$\dim \mathscr{V}_{\boldsymbol{d},\boldsymbol{r}} \leq \min\bigg(\underbrace{\sum_{\ell=1}^{L} d_{\ell} d_{\ell-1} - \sum_{\ell=1}^{L-1} d_{\ell}}_{\text{degrees of freedom}}, \; \underbrace{\dim \left( (\mathscr{H}_{d_0,r_{total}})^{\times d_L} \right)}_{\text{output space dimension}} \bigg).$$

If the bound is reached, we say that the neurovariety has *expected dimension*. Moreover, if the right bound is reached, i.e.,

$$\dim \mathscr{V}_{\boldsymbol{d},\boldsymbol{r}} = \dim((\mathscr{H}_{d,r})^{\times d_L}),$$

the hPNN is *expressive*, and the neurovariety  $\mathscr{V}_{d,r}$  is said to be *thick* [7]. As a consequence, any polynomial map of degree  $r_{total}$  in  $\mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$  can be represented as an hPNN with layer widths  $(d_0, 2d_1, \ldots, 2d_{L-1}, d_L)$  and activation degrees  $r_1 = r_2 = \cdots = r_{L-1}$  [7, Proposition 5].

The left bound  $(\sum_{\ell=1}^L d_\ell d_{\ell-1} - \sum_{\ell=1}^{L-1} d_\ell)$  follows from the equivalences Lemma 4 and defines the number of effective parameters of the representation. This bound is tightly linked with identifiability, as shown in the following well-known lemma.

**Lemma D.1** (Dimension and number of decompositions). The dimension of  $\mathcal{V}_{d,r}$  satisfies

$$\dim \mathscr{V}_{\boldsymbol{d},\boldsymbol{r}} = \sum_{\ell=1}^L d_\ell d_{\ell-1} - \sum_{\ell=1}^{L-1} d_\ell$$

if and only if the map  $\Psi_{d,r}$  is finite-to-one, that is, the generic fiber (i.e. preimage  $\Psi_{d,r}^{-1}(\Psi_{d,r}(w))$  for general w) contains a finite number of the equivalence classes defined in Lemma 4.

Our identifiability results for hPNNs are finite-to-one, and thus lead to the following immediate corollary on the expected dimension of  $\mathcal{V}_{d,r}$ :

**Corollary D.2** (Identifiability implies expected dimension). *If the architecture* (d, r) *is identifiable according to Definition 8, then the neurovariety*  $\mathcal{V}_{d,r}$  *has expected dimension.* 

**Example D.3.** *In Example 9*,  $\dim(\mathscr{V}_{d,r}) = 6 + 4 - 2 = 8$ .

# E Truncation of PNNs with bias terms

In this appendix, we describe an alternative (to homogenization) approach to prove the identifiability of the weights  $W_\ell$  of  $\mathrm{PNN}_{d,r}[(w,b)]$  based on *truncation*. The key idea is that the truncation of a PNN is an hPNN, which allow one to leverage the uniqueness results for hPNNs. However, we note that unlike homogeneization, truncation does not by itself guarantees the identifiability of the bias terms  $b_\ell$ .

For truncation, we use leading terms of polynomials, i.e. for  $p \in \mathscr{P}_{d,r}$  we define  $\operatorname{lt}\{p\} \in \mathscr{H}_{d,r}$  the homogeneous polynomial consisting of degree-r terms of p:

**Example E.1.** For a bivariate polynomial  $p \in \mathcal{P}_{2,2}$  given by

$$p(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2 + ex_1 + fx_2 + g,$$

its truncation  $q = lt\{p\} \in \mathcal{H}_{2,2}$  becomes

$$q(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2.$$

In fact  $\mathrm{lt}\{\cdot\}$  is an orthogonal projection  $\mathscr{P}_{d,r} \to \mathscr{H}_{d,r}$ ; we also apply  $\mathrm{lt}\{\cdot\}$  to vector polynomials coordinate-wise. Then, PNNs with biases can be treated using the following lemma.

**Lemma E.2.** Let  $p = PNN_{d,r}[(w, b)]$  be a PNN with bias terms. Then its truncation is the hPNN with the same weight matrices

$$lt\{p\} = hPNN_{d,r}[w].$$

*Proof.* The statement follows from the fact that  $lt\{(q(\boldsymbol{x}))^r\} = lt\{(q(\boldsymbol{x}))\}^r$ . Indeed, this implies  $lt\{(\langle \boldsymbol{v}, \boldsymbol{x} \rangle + \boldsymbol{c})^r\} = (\langle \boldsymbol{v}, \boldsymbol{x} \rangle)^r$ , which can be applied recursively to  $PNN_{\boldsymbol{d}, \boldsymbol{r}}[(\boldsymbol{w}, \boldsymbol{b})]$ .

**Example E.3.** Consider a 2-layer PNN

$$f(x) = W_2 \rho_{r_1} (W_1 x + b_1) + b_2. \tag{27}$$

Then its truncation is given by

$$\operatorname{lt}\{f\}(\boldsymbol{x}) = \boldsymbol{W}_2 \rho_{r_1}(\boldsymbol{W}_1 \boldsymbol{x}).$$

This idea is well-known and in fact was used in [44] to analyze identifiability of a 2-layer network with arbitrary polynomial activations.

**Remark E.4.** Thanks to Lemma E.2, the identifiability results obtained for hPNNs can be directly applied. Indeed, we obtain identifiability of weights, under the same assumptions as listed for the hPNN case. However, this does not guarantee identifiability of biases, which was achieved using homogeneization.

# **NeurIPS Paper Checklist**

## 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We provide a clear description of our contributions in Section 1.1 in the introduction.

#### Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals
  are not attained by the paper.

## 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Our paper is of a theoretical nature, we outline all the assumptions used to derive the results very clearly.

## Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

## 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We clearly state the assumptions used in each theoretical result, and complete proofs are provided in an appendix in the supplemental material.

## Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and crossreferenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

# 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: Our paper does not include experiments.

#### Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
- (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

# 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Our paper is theoretical and does not include experiments, thus it does not require data or code.

## Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: Our paper is theoretical and does not include experiments.

# Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

# 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Our paper is theoretical and does not include experiments.

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: Our paper is theoretical and does not include experiments.

#### Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster. or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Our research adhered to every aspect of the NeurIPS Code of Ethics.

## Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

# 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our paper is theoretical in nature (we study the identifiability/uniqueness of neural networks). Thus, our results advance the understanding of the behavior of neural networks. We highlighted the potential positive impacts of our identifiability results on the interpretability of neural networks in the paper.

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

# 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper poses no such risks.

## Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

# 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Our paper does not use existing assets.

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

# 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not release new assets.

# Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

# 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects. Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

# 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects. Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

# 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in our research does not involve LLMs as any important, original, or non-standard components.

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.