

A roadmap to the appendices⁴

The appendices of the paper contain background on tensor decompositions and neurovarieties, the proofs of the technical results, as well as a discussion on the changes between the originally submitted and final version of the paper. They are organized as follows:

- Appendix A presents background on neurovarieties for homogeneous PNNs. This is a crucial part for understanding the link between finite identifiability of an hPNN, the dimension of its neurovariety and the rank of the Jacobian of its parametrization map.
- Appendix B contains the main technical tools used in the proof the localization theorem and follows the structure of Section 4. In particular, it presents the proofs of necessary conditions for uniqueness (Section 4.1), background on tensor decompositions and Kruskal-based sufficient conditions for the identifiability of 2-layer hPNNs (Section 4.2).
- Appendix C presents the proof of the localization theorem (Theorem 11) and its consequences for several hPNN architectures, as well as some supporting technical results.
- Appendix D presents the proofs for the case of PNNs with biases. Appendix D.3 discusses the idea of *truncation*, an alternative approach to tackle the PNNs with biases.
- Appendix E discusses necessary and sufficient conditions for the identifiability of hPNNs, as well as changes between the originally submitted and the final version of the paper which were done to correct a mistake in the proof of one of the main results.

A 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, as explained in this appendix.

A.1 Neurovarieties and dimension

An hPNN architecture (\mathbf{d}, \mathbf{r}) defines a map $\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot]$ from the weight tuple $\mathbf{w} = (\mathbf{W}_1, \dots, \mathbf{W}_L)$ to a (polynomial) function space \mathcal{H} :

$$\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot] : \mathbf{w} \mapsto \text{hPNN}_{\mathbf{d}, \mathbf{r}}[\mathbf{w}] \\ \mathbb{R}^{\sum_{\ell} d_{\ell} d_{\ell-1}} \rightarrow \mathcal{H}.$$

The space \mathcal{H} is the space of length- d_L vectors of homogeneous polynomials of degree $r_{\text{total}} = r_1 r_2 \dots r_{L-1}$ in d_0 variables:

$$\mathcal{H} := (\mathcal{H}_{d_0, r_{\text{total}}})^{\times d_L};$$

thus \mathcal{H} is a finite-dimensional vector space of dimension

$$N = \dim(\mathcal{H}) = d_L \binom{d_0 + r_{\text{total}} - 1}{r_{\text{total}}},$$

which follows from the fact that $\dim(\mathcal{H}_{d, r}) = \binom{d+r-1}{r}$.

The key observation is that $\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot]$ is a *polynomial-in-the-parameters* map, which has important implication on the space of networks with a given architecture. The image $\text{Im}(\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot])$, called a *neuromanifold*, is a semi-algebraic set⁵. The properties of $\text{Im}(\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot])$ are tightly linked to the properties of the *neurovariety* $\mathcal{V}_{\mathbf{d}, \mathbf{r}}$ defined as the closure of $\text{Im}(\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot])$ in the Zariski topology, i.e., the smallest algebraic set⁶ containing $\text{Im}(\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot])$. The key property is the dimension of the neurovariety⁷ which is equal to the dimension of the neurovariety [89, Prop. 2.8.2].

The properties of neurovarieties depend on the field (i.e., results can differ between \mathbb{R} or \mathbb{C}), and we focus on the real case. However, most of the results can be translated to the complex case as well. We

⁴The appendices have been reorganized and reworked for better readability.

⁵[89, Def. 2.1.1]: a set cut out by polynomial equations and inequalities.

⁶[89, Def. 2.1.4]: a set cut out by polynomial equations.

⁷roughly defined as the dimension of the tangent space at general point, see [89, §2.8] for more details.

mostly follow [90, Section 4], and an overview on semialgebraic sets can be also found in [91] (see [89] for a detailed account).

The following upper bound on $\dim \mathcal{V}_{\mathbf{d}, \mathbf{r}}$ the bound was presented in [7]:

$$\dim \mathcal{V}_{\mathbf{d}, \mathbf{r}} \leq \min \left(\underbrace{\sum_{\ell=1}^L d_\ell d_{\ell-1} - \sum_{\ell=1}^{L-1} d_\ell}_{\text{degrees of freedom}}, \underbrace{\dim \mathcal{H}}_{\text{output space dimension}} \right). \quad (8)$$

If the bound in (8) is reached, we say that the neurovariety has *expected dimension*. There are two fundamentally different cases when the expected dimension is reached.

Expressive case. If the right bound is reached, i.e., the neurovariety:

$$\dim \mathcal{V}_{\mathbf{d}, \mathbf{r}} = \dim (\mathcal{H}) = d_L \binom{d_0 + r_{\text{total}} - 1}{r_{\text{total}}},$$

the hPNN is *expressive*, and the neurovariety $\mathcal{V}_{\mathbf{d}, \mathbf{r}}$ is said to be *thick* [7], as it fills the whole function space \mathcal{H} (and thus the neuromanifold is of positive Lebesgue measure). In particular, this implies that (see [7, Proposition 5]) any homogeneous polynomial vector from \mathcal{H} (i.e., of degree r_{total} with d_0 inputs and d_L outputs, with degrees fixed as $r_1 = r_2 = \dots = r_{L-1}$) can be represented as an hPNN with layer widths $(d_0, 2d_1, \dots, 2d_{L-1}, d_L)$ and the same activation degrees.

Identifiable case. The left bound $(\sum_{\ell=1}^L d_\ell d_{\ell-1} - \sum_{\ell=1}^{L-1} d_\ell)$ follows from the presence of equivalences defined in Lemma 4 (i.e., the size of the vector \mathbf{w} minus the number of independent rescalings) and defines the number of effective parameters of the representation (this is explained in the following subsections). Moreover, the left bound is reached if and only if the hPNN architecture is finitely identifiable:

Proposition 10 *The architecture $\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot]$ is finitely identifiable if and only if the dimension of $\mathcal{V}_{\mathbf{d}, \mathbf{r}}$ is equal to the effective number of parameters, i.e., $\dim \mathcal{V}_{\mathbf{d}, \mathbf{r}} = \sum_{\ell=1}^L d_\ell d_{\ell-1} - \sum_{\ell=1}^{L-1} d_\ell$. In such case, $\mathcal{V}_{\mathbf{d}, \mathbf{r}}$ is said to be **nondefective**. Equivalently, the rank of the Jacobian of the map $\text{hPNN}_{\mathbf{d}, \mathbf{r}}[\cdot]$ is maximal and equal to $\sum_{\ell=1}^L d_\ell d_{\ell-1} - \sum_{\ell=1}^{L-1} d_\ell$ at a general parameter \mathbf{w} .*

Proposition 10 is central to the proof of the main results of paper. The proof Proposition 10 relies on properties of fibers of polynomial maps and is reviewed in the next subsection, together with the Jacobian of the parameterization.

A.2 Polynomial maps and fiber dimension

We recall some key facts on the polynomial maps and their images. We begin by highlighting the link between dimensions of semialgebraic sets and the Jacobian of the polynomial maps.

Lemma A.1. *Let $\varphi : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be a polynomial map, and denote by $J_\varphi(\boldsymbol{\theta})$ the $n \times m$ Jacobian matrix. Let*

$$r_0 := \max_{\boldsymbol{\theta}} \text{rank}\{J_\varphi(\boldsymbol{\theta})\}.$$

Then we have that:

1. $\text{rank}\{J_\varphi(\boldsymbol{\theta})\} = r_0$ for generic $\boldsymbol{\theta}$ (i.e., for all $\boldsymbol{\theta} \in \mathbb{R}^m$ except a set of Lebesgue measure zero, where the rank of the Jacobian is strictly less than r_0).
2. r_0 is equal to the dimension of $\text{Im}(\varphi)$ and its (Zariski) closure:

$$r_0 = \dim(\text{Im}(\varphi)) = \dim(\overline{\text{Im}(\varphi)}).$$

The proof of Lemma A.1 is given in [90, Theorem 4.7] and the preceding paragraph (in [90] r_0 is called *generic rank* of the parameterization φ). It mainly follows from semicontinuity of the rank of a matrix.

Remark A.2 (On genericity). *Due to the algebraic structure, the genericity statement in Lemma A.1 is much stronger: in fact, the set of points θ where $\text{rank}\{J_\varphi(\theta)\} \neq r_0$ is a semialgebraic subset of \mathbb{R}^m of dimension strictly less than m . The same holds for all generic statements and definitions in the paper (such as finite identifiability, global identifiability, etc.), see the definition of genericity in [90, Definition 4.1].*

Remark A.3. *The right bound in (8) follows essentially from Lemma A.1: indeed, in the case $\varphi(\cdot) = \text{hPNN}_{d,r}[\cdot]$, $\text{rank}\{J_\varphi\}$ does not exceed the dimension of the ambient space of φ (equal to $\dim(\mathcal{H})$).*

The following lemma is key for linking finite identifiability to the dimension of the neurovariety.

Lemma A.4 (Fiber dimension). *Let $\varphi : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be a polynomial map, so that $r_0 = \dim(\text{Im}(\varphi))$. Then the dimension of its generic fiber is equal to $m - r_0$, that is, for generic $\theta \in \mathbb{R}^m$, the preimage $\varphi^{-1}(\varphi(\theta))$ is a semialgebraic set with*

$$\dim \varphi^{-1}(\varphi(\theta)) = m - r_0.$$

Lemma A.4 is well known to specialists, but in the literature it is mostly formulated for the complex case (see [90, Theorem 4.7]). For the real field it follows from [90, Theorem 4.9].

A particular case is when $r_0 = m$, in which case Lemma A.4 implies finiteness of the fiber:

Corollary A.5. *The following two statements are equivalent:*

- *For general $\theta \in \mathbb{R}^m$, $\text{rank}\{J_\varphi(\theta)\} = m$;*
- *For general $\theta \in \mathbb{R}^m$, the fiber (i.e., the preimage $\varphi^{-1}(\varphi(\theta))$) consists of a finite number of points.*

Corollary A.5 follows simply from the fact that 0-dimensional semialgebraic sets are collections of a finite number of points.

Finally we make the following remark that is very commonly used.

Corollary A.6. *If there exists θ_0 such that $\text{rank}\{J_\varphi(\theta_0)\} = m$, then the $\text{rank}\{J_\varphi(\theta)\} = m$ for generic θ .*

Proof. This follows from Lemma A.1 as r_0 from Lemma A.1 is equal to m . \square

Remark A.7. *Corollary A.6 implies that finding a single point with full column rank Jacobian implies finiteness of the generic fiber.*

A.2.1 The case of neurovarieties

The first implication of Lemma A.4 is the left upper bound in (8). It is based on the following lemma from [7], for which we provide a short proof for completeness.

Lemma A.8 ([7, Lemma 13]). *For a general parameter $w = (\mathbf{W}_1, \dots, \mathbf{W}_L)$, the set of equivalent hPNN representations in Lemma 4 is semialgebraic and of dimension $\sum_{\ell=1}^{L-1} d_\ell$.*

Proof. First, note that the set of equivalent representation is of dimension at most $\sum_{\ell=1}^{L-1} d_\ell$ (by the number of parameters). Consider a general w , so that the first column of each \mathbf{W}_ℓ , for $\ell = 1, \dots, L-1$, equal to $\mathbf{v}_\ell \in \mathbb{R}^{d_\ell}$, does not have zero elements. Now take any collection of vectors $\tilde{\mathbf{v}}_1 \in \mathbb{R}^{d_1}, \dots, \tilde{\mathbf{v}}_{L-1} \in \mathbb{R}^{d_{L-1}}$ having elementwise the same signs as \mathbf{v}_ℓ . Then there exist matrices \mathbf{D}_ℓ so that the equivalent weight matrices $\tilde{\mathbf{W}}_\ell = \tilde{\mathbf{D}}_\ell \mathbf{W}_\ell \tilde{\mathbf{D}}_{\ell-1}^{-r_{\ell-1}}$ have $\tilde{\mathbf{v}}_\ell$ exactly as their first columns. Thus the set of equivalent representations is exactly of dimension $\sum_{\ell=1}^{L-1} d_\ell$. \square

Remark A.9. *The left upper bound in (8) simply follows from Lemma A.8 (as written in [7, Lemma 13]): indeed, the dimension of the fiber of $\text{hPNN}_{d,r}[\cdot]$ must be at least $\sum_{\ell=1}^{L-1} d_\ell$. This implies, by Lemma A.4,*

$$\text{rank}\{J_\varphi(\theta)\} \leq \sum_{\ell=1}^L d_{\ell-1} d_\ell - \sum_{\ell=1}^{L-1} d_\ell, \quad (9)$$

which is exactly the right dimension bound in (8) by Lemma A.1.

Note that Proposition 10 will exactly consider the case when the equality is reached in (9) for generic θ . Similarly to Corollary A.6, the following corollary of Lemma A.1 implies that for the case of neurovarieties it suffices to find a single set of parameters w where the Jacobian of the parameterization is of maximal rank to guarantee finite identifiability of hPNN architecture. This will be used in the proofs to give a *certificate* of finite identifiability

Corollary A.10. *If there exists a particular point θ_0 such that equality is achieved in (9), then the equality in (9) is achieved for generic θ .*

Proof. Since there exists such a θ_0 , then the r_0 defined in Lemma A.1 satisfies

$$r_0 \geq \sum_{\ell=1}^L d_{\ell-1} d_{\ell} - \sum_{\ell=1}^{L-1} d_{\ell}. \quad (10)$$

But from (9), r_0 must be bounded from above by the same number. Therefore the equality for r_0 is achieved in (10). \square

A.3 Proof of the proposition

Proof of Proposition 10. We denote $\varphi(\cdot) = \text{hPNN}_{d,r}[\cdot]$ for simplicity (so that $m = \sum_{\ell=1}^L d_{\ell} d_{\ell-1}$ and $n = \dim \mathcal{H}$) and consider separately the “only if” (\Rightarrow) and “if” (\Leftarrow) parts.

\Rightarrow Assume that for a generic w the fiber $\varphi^{-1}(\varphi(w))$ consists of finite number of equivalence classes, thus it is a finite union of non-intersecting semialgebraic subsets of dimension $\sum_{\ell=1}^{L-1} d_{\ell}$. Therefore, by [89, Theorem 2.8.5] the whole fiber $\varphi^{-1}(\varphi(w))$ has the dimension equal to $\sum_{\ell=1}^{L-1} d_{\ell}$ as well, hence $\dim \mathcal{V}_{d,r} = \sum_{\ell=1}^L d_{\ell} d_{\ell-1} - \sum_{\ell=1}^{L-1} d_{\ell}$.

\Leftarrow The proof follows a similar argument as in the proof of [90, Theorem 4.9]. We consider a (Zariski open) subset of parameters without zero values $\mathcal{U} = (\mathbb{R} \setminus \{0\})^m$. It can be shown that the preimage of the image of its complement $\mathcal{Z} := \varphi^{-1}(\varphi(\mathbb{R}^m \setminus \mathcal{U}))$ is a (semialgebraic) set of measure zero. Therefore for the set $\mathcal{U}' := \mathcal{U} \setminus \mathcal{Z}$ the preimage of the image is contained in \mathcal{U} :

$$\varphi^{-1}(\varphi(\mathcal{U}')) \subset \mathcal{U}.$$

Note that any $w \in \mathcal{U}$ can be brought (by diagonal scaling and permutation) to the equivalent form:

$$W_{\ell} = \begin{bmatrix} 1 & \dots & 1 \\ \overline{W}_{\ell} \end{bmatrix}, \quad \overline{W}_{\ell} \in \mathbb{R}^{(d_{\ell}-1) \times d_{\ell-1}} \quad (11)$$

for all $\ell = 2, \dots, L$ where the reduced \overline{W}_{ℓ} parameterize the classes of equivalent parameters in \mathcal{U} up to permutation. Now denote $\overline{w} = (W_1, \overline{W}_2, \dots, \overline{W}_L)$ and

$$\psi : \overline{w} \mapsto \text{hPNN}_{d,r}[w].$$

If we can guarantee that the generic fiber of ψ is finite, then this will imply that on \mathcal{U}' , the fiber of the map φ contains finitely many equivalence classes. For this we note that the Jacobian of ψ is just a submatrix of the Jacobian of φ with exactly $m - \sum_{\ell=1}^{L-1} d_{\ell}$ columns. We will show that it is full rank at a generic point \overline{w} .

Consider the following map

$$\xi : (W_1, \overline{W}_2, \dots, \overline{W}_L, D_1, \dots, D_L) \mapsto (W_1, \widetilde{W}_2, \dots, \widetilde{W}_L)$$

defined as

$$\widetilde{W}_{\ell} = D_{\ell} \begin{bmatrix} 1 & \dots & 1 \\ \overline{W}_{\ell} \end{bmatrix} D_{\ell}^{-r_{\ell-1}}$$

for $\ell = 2, \dots, L$ (with the convention that $D_L = I_{d_L}$).

Consider a particular \overline{w}_0 constructed as above (by normalization of a $w \in \overline{\mathcal{U}}$). Then for a neighbourhood $\overline{\mathcal{U}}$ of \overline{w}_0 and a neighbourhood \mathcal{V} of $(I_{d_1}, \dots, I_{d_{L-1}})$, the map ξ is a diffeomorphism from $\overline{\mathcal{U}} \times \mathcal{V}$ to an open neighbourhood of the corresponding w_0 (defined by (11)).

Consider the composition $\varphi \circ \xi$. Then at the point $(\mathbf{W}_1, \overline{\mathbf{W}}_2, \dots, \overline{\mathbf{W}}_L, \mathbf{I}_{d_1}, \dots, \mathbf{I}_{d_{L-1}})$, we have that (i) the derivatives with respect to \mathbf{D}_ℓ at identity matrices are zero and (ii) the Jacobian of $\varphi \circ \xi$ with respect to $\overline{\mathbf{w}}$ coincides with the Jacobian of ψ , hence it must have full column rank $(m - \sum_{\ell=1}^{L-1} d_\ell)$ equal to the dimension of the neurovariety. Hence, the fiber of ψ is finite, which implies finite identifiability of φ . \square

B Main tools for the proof

This appendix contains the main technical tools used in the proof the localization theorem. It is organized in three subsections, following the same structure as in Section 4:

- Appendix B.1 presents the proofs of necessary conditions for uniqueness corresponding to Section 4.1 in the main body of the paper;
- Appendix B.2 presents background on tensor decompositions and the proof of Proposition 34 from the main body of the paper, which shows the link between 2-layer hPNNs and partially symmetric tensors;
- Appendix B.3 presents Kruskal-based sufficient conditions for the identifiability of 2-layer hPNNs (Propositions 35 and 12 in the main paper).

B.1 Necessary conditions for uniqueness

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.

Lemma 30. *Let $\mathbf{p} = \text{hPNN}_r[\mathbf{w}]$ be an hPNN of format $(d_0, \dots, d_\ell, \dots, d_L)$. Then for any ℓ there exists an infinite number of representations of hPNNs $\mathbf{p} = \text{hPNN}_r[\mathbf{w}]$ 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 $(\mathbf{W}_0, \dots, \mathbf{W}_L)$ the weight matrices associated with the representation of format $(d_0, \dots, d_\ell, \dots, d_L)$ of the hPNN $\mathbf{p} = \text{hPNN}_r[\mathbf{w}]$. By assumptions on the dimensions, the two matrices $\mathbf{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and $\mathbf{W}_{\ell+1} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ read

$$\begin{aligned} \mathbf{W}_{\ell+1} &= [\mathbf{w}_1 \quad \dots \quad \mathbf{w}_{d_\ell}], \text{ where, for each } i, \mathbf{w}_i \in \mathbb{R}^{d_{\ell+1}}, \\ \mathbf{W}_\ell &= [\mathbf{v}_1 \quad \dots \quad \mathbf{v}_{d_\ell}]^\top, \text{ where, for each } i, \mathbf{v}_i \in \mathbb{R}^{d_{\ell-1}}. \end{aligned}$$

Without loss of generality, let us assume that \mathbf{w}_i are nonzero, and set

$$\widetilde{\mathbf{W}}_\ell = [\mathbf{0} \quad \mathbf{v}_1 \quad \dots \quad \mathbf{v}_{d_\ell}]^\top \in \mathbb{R}^{(d_\ell+1) \times d_{\ell-1}},$$

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

$$\widetilde{\mathbf{W}}_{\ell+1}^{(\mathbf{u})} = [\mathbf{u} \quad \mathbf{w}_1 \quad \dots \quad \mathbf{w}_{d_\ell}] \in \mathbb{R}^{d_{\ell+1} \times (d_\ell+1)}.$$

Then, we have that for any choice of \mathbf{u} and for any \mathbf{z} ,

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

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

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

of format $(d_0, \dots, d_\ell + 1, \dots, d_L)$ for the hPNN $\mathbf{p} = \text{hPNN}_r[\mathbf{w}]$. \square

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 $\mathbf{p} = \text{hPNN}_r[\mathbf{w}]$ be a unique L -layers decomposition. Consider the vector output at any ℓ -th internal level $\ell < L$ after the activations*

$$\mathbf{q}_\ell(\mathbf{x}) = \rho_{r_\ell} \circ \mathbf{W}_\ell \circ \dots \circ \rho_{r_1} \circ \mathbf{W}_1(\mathbf{x}).$$

Then the elements $\mathbf{q}_\ell(\mathbf{x}) = [q_{\ell,1}(\mathbf{x}) \ \dots \ q_{\ell,d_\ell}(\mathbf{x})]^\top$ are linearly independent polynomials.

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

$$\mathbf{p} = \mathbf{W}_L \circ \rho_{r_{L-1}} \circ \dots \circ \rho_{r_{\ell+1}} \circ \mathbf{W}_{\ell+1} \mathbf{B} \begin{bmatrix} q_{\ell,1}(\mathbf{x}) \\ \vdots \\ q_{\ell,d_\ell-1}(\mathbf{x}) \end{bmatrix},$$

i.e., the hPNN \mathbf{p} admits a representation of size $\mathbf{d} = (d_0, \dots, d_\ell - 1, \dots, d_L)$ with parameters $(\mathbf{W}_1, \dots, \mathbf{W}_{\ell+1} \mathbf{B}, \dots, \mathbf{W}_L)$. Therefore, by Lemma 30 its original representation is not unique, which is a contradiction. \square

Using Lemma 30 and Lemma 31, we can prove the conditions on the Kruskal ranks of weight matrices that are necessary for uniqueness. These conditions are based on the notion of Kruskal rank which we recall from [15].

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

Note that the following two cases of particular interest also have simple equivalent interpretations:

- $\text{krank}\{\mathbf{A}\} \geq 1$ is equivalent to saying that matrix \mathbf{A} has no zero columns;
- $\text{krank}\{\mathbf{A}\} \geq 2$ is equivalent to saying that no pair of the columns of matrix \mathbf{A} are collinear.

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

$$\text{krank}\{\mathbf{W}_\ell^\top\} \geq 2, \quad \text{krank}\{\mathbf{W}_{\ell+1}\} \geq 1,$$

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

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

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

B.2 Background on tensors

In this appendix, we first present a background on tensors and the CP tensor decomposition, and demonstrate the link between hPNNs and the partially symmetric CPD (Proposition 34 in the main paper).

B.2.1 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., \mathbf{a} . Matrices (tensors of order two) are denoted by boldface capital letters, e.g., \mathbf{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 , $\mathcal{T} \in \mathbb{R}^{m_1 \times \dots \times m_s}$ is the matrix $\mathbf{T}^{(p)}$ of size $m_r \times (m_1 m_2 \dots m_{r-1} m_{r+1} \dots m_s)$ defined as

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

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

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

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

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

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

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

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

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

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

$$[\mathcal{T} \bullet_r \mathbf{A}]_{i_1, \dots, i_{r-1}, j, i_{r+1}, \dots, i_s} = \sum_{i_r=1}^{m_r} \mathcal{T}_{i_1, \dots, i_s} \mathbf{A}_{j, i_r}.$$

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

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

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

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

where $\mathbf{A}^{(p)} = [\mathbf{a}_1^{(p)} \dots \mathbf{a}_R^{(p)}] \in \mathbb{R}^{m_p \times R}$.

When \mathcal{T} is partially symmetric along the modes $(p+1, \dots, s)$, for $p < s$, its CPD satisfies $\mathbf{A}^{(p+1)} = \mathbf{A}^{(p+2)} = \dots = \mathbf{A}^{(s)}$. 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 $\mathcal{T} \in \mathbb{R}^{m \times m \times \dots \times m}$ is defined as

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

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 $1 \times m \times \dots \times m$, as

$$\mathcal{T} = \llbracket \mathbf{u}, \mathbf{A}, \dots, \mathbf{A} \rrbracket,$$

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

B.2.2 Link between hPNNs and partially symmetric tensors

Recall that $\mathcal{H}_{d_0,r}$ denotes the space of d_0 -variate homogeneous 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 $\mathcal{F} \in \mathbb{R}^{d_2 \times d_0 \times \dots \times d_0}$ and polynomial vectors $\mathbf{f} \in (\mathcal{H}_{d_0,r})^{\times d_2}$, which can be written as*

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

with $\mathbf{F}^{(1)} \in \mathbb{R}^{d_2 \times d_0^r}$ the first unfolding of \mathcal{F} . Under this mapping, the partially symmetric CPD

$$\mathcal{F} = \llbracket \mathbf{W}_2, \mathbf{W}_1^\top, \dots, \mathbf{W}_1^\top \rrbracket$$

is mapped to hPNN $\mathbf{W}_2 \rho_r(\mathbf{W}_1 \mathbf{x})$. Thus, uniqueness of $\text{hPNN}_{(d_0, d_1, d_2), r}[\mathbf{W}_1, \mathbf{W}_2]$ is equivalent to uniqueness of the partially symmetric CPD of \mathcal{F} .

Proof. We distinguish the two cases, $d_2 = 1$ and $d_2 \geq 2$. We begin the proof by the more general case $d_2 \geq 2$.

Case $d_2 \geq 2$. Denoting by $\mathbf{u}_i \in \mathbb{R}^{d_2}$ the i -th column of \mathbf{W}_2 and $\mathbf{v}_i \in \mathbb{R}^{d_0}$ the i -th row of \mathbf{W}_1 , the relationship between the 2-layer hPNN and tensor \mathcal{F} can be written explicitly as

$$\begin{aligned} \mathbf{f}(\mathbf{x}) &= \mathbf{W}_2 \rho_r(\mathbf{W}_1 \mathbf{x}) \\ &= \sum_{i=1}^{d_1} \mathbf{u}_i (\mathbf{v}_i^\top \mathbf{x})^r \\ &= \sum_{i=1}^{d_1} \mathbf{u}_i (\mathbf{v}_i^{\otimes r})^\top \mathbf{x}^{\otimes r} \\ &= \underbrace{\mathbf{W}_2 (\mathbf{W}_1^\top \odot \dots \odot \mathbf{W}_1^\top)^\top}_{=\mathbf{F}^{(1)}} \mathbf{x}^{\otimes r}, \end{aligned}$$

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 $d_2 = 1$. When $d_2 = 1$, the columns of $\mathbf{W}_2 \in \mathbb{R}^{1 \times d_1}$ are scalar values $u_i \in \mathbb{R}$, $i = 1, \dots, d_1$. In this case, $(\mathbf{W}_1^\top \odot \dots \odot \mathbf{W}_1^\top) \mathbf{W}_2^\top$ becomes equivalent to the vectorization of \mathcal{F} , which is a fully symmetric tensor of order r with factors \mathbf{W}_1^\top and coefficients $[\mathbf{W}_2]_{1,i}$, $i = 1, \dots, d_1$. \square

B.3 Kruskal-based conditions for the uniqueness and identifiability of 2-layer networks

B.3.1 Sufficient conditions for uniqueness

The direct links between 2-layer ($L = 2$) hPNNs and partially symmetric CPDs in Proposition 34 allows us to obtain sufficient conditions for their uniqueness by means of Kruskal-based uniqueness results for the CPD, which we recall in the following lemma.

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

$$\sum_{i=1}^s \text{krank}\{\mathbf{A}^{(i)}\} \geq 2R + (s-1). \quad (12)$$

Then the CP decomposition of \mathcal{T} is unique up to permutation and scaling ambiguities, that is, for any alternative CPD $\mathcal{T} = \llbracket \tilde{\mathbf{A}}^{(1)}, \tilde{\mathbf{A}}^{(2)}, \dots, \tilde{\mathbf{A}}^{(s)} \rrbracket$, there exist a permutation matrix Π and invertible diagonal matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_s$ such that

$$\tilde{\mathbf{A}}^{(i)} = \mathbf{A}^{(i)} \Pi \Lambda_i,$$

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

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

Proposition 35. *Let $\mathbf{p}_w(\mathbf{x}) = \mathbf{W}_2 \rho_{r_1}(\mathbf{W}_1 \mathbf{x})$ be a 2-layer hPNN with $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d_0}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times d_1}$ and layer sizes (d_0, d_1, d_2) satisfying $d_0, d_1 \geq 2, d_2 \geq 1$. Assume that $r \geq 2$, $\text{krank}\{\mathbf{W}_2\} \geq 1$, $\text{krank}\{\mathbf{W}_1^\top\} \geq 2$ and that:*

$$r \geq \frac{2d_1 - \text{krank}\{\mathbf{W}_2\}}{\text{krank}\{\mathbf{W}_1^\top\} - 1},$$

then the 2-layer hPNN $\mathbf{p}_w(\mathbf{x})$ 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 $\mathbf{p}_w(\mathbf{x})$ is in one-to-one correspondence with the order $r + 1$ partially symmetric tensor

$$\mathcal{F} = \llbracket \mathbf{W}_2, \mathbf{W}_1^\top, \dots, \mathbf{W}_1^\top \rrbracket, \quad (13)$$

thus, the uniqueness of $\mathbf{p}_w(\mathbf{x})$ is equivalent to that of the CP-decomposition of \mathcal{F} in (13). From [82, Theorem 3], the rank- d CP decomposition of \mathcal{T} is unique provided that

$$\text{krank}\{\mathbf{W}_2\} + r \text{krank}\{\mathbf{W}_1^\top\} \geq 2d_1 + r.$$

By noting that $\text{krank}\{\mathbf{W}_1^\top\} > 1$ and rearranging the terms, we obtain the desired result. \square

Note that for the case of $d_0 \geq 2$ (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., $d_2 = 1$), it requires $r \geq 3$.

B.3.2 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 $d_0, d_1 \geq 2, d_2 \geq 1$ be the layer widths and $r \geq 2$ such that*

$$r \geq \frac{2d_1 - \min(d_1, d_2)}{\min(d_1, d_0) - 1}.$$

Then the 2-layer hPNN with architecture $((d_0, d_1, d_2), (r))$ is globally identifiable.

Proof of Proposition 12. For general matrices $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d_0}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_2 \times d_1}$, we have

$$\begin{aligned} \text{krank}\{\mathbf{W}_1^\top\} &= \min(d_0, d_1), \\ \text{krank}\{\mathbf{W}_2\} &= \min(d_2, d_1). \end{aligned}$$

Moreover, $d_0, d_1 \geq 2, d_2 \geq 1$ implies that generically $\text{krank}\{\mathbf{W}_1^\top\} \geq 2$ and $\text{krank}\{\mathbf{W}_2\} \geq 1$. 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 $((d_0, d_1, d_2), (r))$ is globally identifiable. \square

C 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.

Roadmap of the proof: The proof of the localization theorem requires some setup. The main idea, as briefly sketched in Section 4.3 of the main paper, is to construct a recursion for Jacobian of the parameter map, and to certify that it has maximal rank (generically). This relies crucially on the properties of the neurovarieties associated to an hPNN as explained in Appendix A, in particular on Proposition 10 and Lemma A.4, which link the finite identifiability of the hPNN to the rank of its Jacobian. The proof of the main result is presented towards the end of this appendix, in Appendix C.7, and proceeds by induction. However, it requires several technical tools which are build in the subsections that precede it.

- Appendix C.1 starts with some preparatory results on the rank of the Jacobian of a 2-layer hPNN, setting the base case.
- Appendix C.2 defines the so-called *last layer map* (i.e., the map that composes a d_0 -variate polynomial with one hPNN layer) and illustrates the structure of its Jacobian by means of a detailed example.
- Appendix C.3 presents a key proposition which establishes a *certificate* to show that the Jacobian of the last layer map has maximal rank, and before proceeding to the proof, illustrates the result with an example.
- Appendix C.4 introduces some additional notation and setup which will be used in the proof of the key proposition.
- Appendix C.5 presents the proof of the key proposition for the special case when the number of input variables d_0 is equal to the number of variables used in the certificate (equal to the smallest bottleneck in the network).
- Appendix C.6 gives the proof of the key proposition in the general case when the number of input neurons d_0 can be larger then the certificate.
- Appendix C.7 contains the proof of the localization theorem.
- Finally, Appendix C.8 presents the proofs for the results concerning the implications of the localization theorem to different hPNN architectures.

Simplifying the notation: In the remainder of this appendix we denote the number of input neurons by m , the number of hidden neurons in the second-to-last layer by d , and the number of output neurons as n . For two-layer networks, we denote the first- and second-layer weight matrices by \mathbf{V} and \mathbf{W} , respectively.

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

Lemma C.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 $n \times m \times \dots \times m$ tensor is unique). Then for general matrices \mathbf{V}, \mathbf{W} the Jacobian of the map $\varphi(\mathbf{V}, \mathbf{W}) = \text{hPNN}_r[(\mathbf{V}, \mathbf{W})]$, given by*

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

has maximal possible rank:

$$\text{rank}\{J_\varphi\} = (m + n - 1)d, \quad (14)$$

and also

$$\text{rank}\{J_\varphi^{(\mathbf{V})}\} = md. \quad (15)$$

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 (\mathbf{V}, \mathbf{W}) with \mathbf{W} given as

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

parameterizes an open subset of the neurovariety (i.e., due to the scaling ambiguity, almost any pair of \mathbf{V} and \mathbf{W} can be reduced to such a form). As shown in the proof of Proposition 10 (specialized to $(\mathbf{W}_1, \mathbf{W}_2) = (\mathbf{V}, \mathbf{W})$), the reduced Jacobian is full column rank:

$$\text{rank}\left\{ \begin{bmatrix} J_\varphi^{(\mathbf{V})} & J_\varphi^{(\overline{\mathbf{W}})} \end{bmatrix} \right\} = md + (n - 1)d,$$

where $J_\varphi^{(\overline{\mathbf{W}})}$ denotes the Jacobian with respect to $\overline{\mathbf{W}}$. Note this implies that all the submatrices are full column rank and, as therefore $J_\varphi^{(\mathbf{V})}$ is full column rank.

□

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

Before giving the elements of the main proof, we provide an example of explicit Jacobian computation for the map $\text{hPNN}_{d,r}[\cdot]$ which will be the guiding example for the proof of identifiability.

Example C.3 (Simplest architecture). Consider example $(m, d, n) = (2, 2, 2)$, $r = 2$, and denote the elements of \mathbf{V} and \mathbf{W} as

$$\mathbf{V} = \begin{bmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} W_{1,1} & W_{1,2} \\ W_{2,1} & W_{2,2} \end{bmatrix}.$$

so the hPNN map $\varphi(\mathbf{V}, \mathbf{W}) = \text{hPNN}_{d,r}[\mathbf{V}, \mathbf{W}]$ is given by

$$\varphi(\mathbf{V}, \mathbf{W}) = \mathbf{w}_1(\alpha_1 x_1 + \beta_1 x_2)^2 + \mathbf{w}_2(\alpha_2 x_1 + \beta_2 x_2)^2,$$

where $\mathbf{w}_1, \mathbf{w}_2$ denote the columns of the matrix \mathbf{W} :

$$\mathbf{w}_1 = \begin{bmatrix} W_{1,1} \\ W_{2,1} \end{bmatrix}, \quad \mathbf{w}_2 = \begin{bmatrix} W_{1,2} \\ W_{2,2} \end{bmatrix}.$$

The image of φ lives in the space of vector polynomials $(\mathcal{H}_{2,2})^{\times 2}$, therefore, the blocks of the Jacobian $J_\varphi^{(\mathbf{V})}$ and $J_\varphi^{(\mathbf{W})}$ are of sizes 6×4 . The matrix $J_\varphi^{(\mathbf{V})}$ has as its columns derivatives with respect to α_j and β_j , for $j \in \{1, 2\}$ which are, respectively:

$$\frac{\partial \varphi}{\partial \alpha_j} = 2\mathbf{w}_j x_1(\alpha_j x_1 + \beta_j x_2), \quad \frac{\partial \varphi}{\partial \beta_j} = 2\mathbf{w}_j x_2(\alpha_j x_1 + \beta_j x_2). \quad (16)$$

Let us choose the canonical basis of $(\mathcal{H}_{2,2})^{\times 2}$ as $\mathbf{e}_i x_1^{2-\ell} x_2^\ell$, $i \in \{1, 2\}$, $\ell \in \{0, 1, 2\}$, where \mathbf{e}_i are unit vectors. Then the block $J_\varphi^{(\mathbf{V})}$ is represented in the matrix form as:

$$J_\varphi^{(\mathbf{V})} = (2) \cdot \begin{array}{c} \begin{matrix} \mathbf{e}_1 x_1^2 \\ \mathbf{e}_1 x_1 x_2 \\ \mathbf{e}_1 x_2^2 \\ \mathbf{e}_2 x_1^2 \\ \mathbf{e}_2 x_1 x_2 \\ \mathbf{e}_2 x_2^2 \end{matrix} \end{array} \begin{bmatrix} \frac{\partial \varphi}{\partial \alpha_1} & \frac{\partial \varphi}{\partial \beta_1} & \frac{\partial \varphi}{\partial \alpha_2} & \frac{\partial \varphi}{\partial \beta_2} \\ W_{1,1}\alpha_1 & 0 & W_{1,2}\alpha_2 & 0 \\ W_{1,1}\beta_1 & W_{1,1}\alpha_1 & W_{1,2}\beta_2 & W_{1,2}\alpha_2 \\ 0 & W_{1,1}\beta_1 & 0 & W_{1,2}\beta_2 \\ W_{2,1}\alpha_1 & 0 & W_{2,2}\alpha_2 & 0 \\ W_{2,1}\beta_1 & W_{2,1}\alpha_1 & W_{2,2}\beta_2 & W_{2,2}\alpha_2 \\ 0 & W_{2,1}\beta_1 & 0 & W_{2,2}\beta_2 \end{bmatrix}$$

The block $J_\varphi^{(\mathbf{W})}$ contains the derivatives with respect to $W_{i,j}$, for $i, j \in \{1, 2\}$, which are:

$$\frac{\partial \varphi}{\partial W_{i,j}} = \mathbf{e}_i(\alpha_j x_1 + \beta_j x_2)^2. \quad (17)$$

In the same monomial basis, the matrix can be expressed as

$$J_\varphi^{(\mathbf{W})} = \begin{array}{c} \begin{matrix} \mathbf{e}_1 x_1^2 \\ \mathbf{e}_1 x_1 x_2 \\ \mathbf{e}_1 x_2^2 \\ \mathbf{e}_2 x_1^2 \\ \mathbf{e}_2 x_1 x_2 \\ \mathbf{e}_2 x_2^2 \end{matrix} \end{array} \begin{bmatrix} \frac{\partial \varphi}{\partial W_{1,1}} & \frac{\partial \varphi}{\partial W_{2,1}} & \frac{\partial \varphi}{\partial \alpha_2} & \frac{\partial \varphi}{\partial \beta_2} \\ \alpha_1^2 & 0 & \alpha_2^2 & 0 \\ 2\alpha_1\beta_1 & 0 & 2\alpha_1\beta_2 & 0 \\ \beta_1^2 & 0 & \beta_2^2 & 0 \\ 0 & \alpha_1^2 & 0 & \alpha_2^2 \\ 0 & 2\alpha_1\beta_1 & 0 & 2\alpha_1\beta_2 \\ 0 & \beta_1^2 & 0 & \beta_2^2 \end{bmatrix}$$

Remark C.4. It is easy to show why (14) and (15) are satisfied for the architecture in Example C.3. For this example, we choose particular \mathbf{V} and \mathbf{W} to be identity matrices, which gives us

$$\begin{bmatrix} J_\varphi^{(\mathbf{V})} & J_\varphi^{(\mathbf{W})} \end{bmatrix} = \left[\begin{array}{cccc|cccc} 2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 1 \end{array} \right].$$

It is easy to see that the left block (matrix $J_\varphi^{(\mathbf{V})}$) has rank 4, and the total matrix has rank $6 = (2 + 2 - 1)2$. Therefore, by Corollaries A.6 and A.10, (14) and (15) are satisfied generically.

We will also need an explicit form of the Jacobian in the general case, which is a generalization of the expression in Example C.3.

Remark C.5. Let (m, d, n) , r , \mathbf{V} and \mathbf{W} be as in Lemma C.1. With some abuse of notation we denote $\mathbf{v}_j \in \mathbb{R}^m$ and $\mathbf{w}_j \in \mathbb{R}^n$

$$\mathbf{V}^\top = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_d], \quad \mathbf{W} = [\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_d],$$

and let $\mathbf{z} = [z_1 \quad \cdots \quad z_m]^\top$ be the input variables. Then the hPNN $\varphi[\cdot] = \text{hPNN}_{d,r}[\cdot]$ has the form

$$\varphi[\mathbf{V}, \mathbf{W}](\mathbf{z}) = \sum_{j=1}^d \mathbf{w}_j (\mathbf{v}_j^\top \mathbf{z})^r. \quad (18)$$

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

$$\frac{\partial \varphi}{\partial W_{i,j}} = \frac{\partial \varphi}{\partial (\mathbf{w}_j)_i} = \mathbf{e}_i (\mathbf{v}_j^\top \mathbf{z})^r, \quad (19)$$

where \mathbf{e}_i is the i -th unit vector in \mathbb{R}^n , and, with respect to elements of \mathbf{V} , we have

$$\frac{\partial \varphi}{\partial V_{j,\ell}} = \frac{\partial \varphi}{\partial (\mathbf{v}_j)_\ell} = r z_\ell \cdot \mathbf{w}_j (\mathbf{v}_j^\top \mathbf{z})^{r-1}. \quad (20)$$

Note that Lemma C.1 concerns the dimensions of linear spaces spanned by the sets of polynomials in (19)–(20). Also, (20) and (19) are generalizations of (16) and (17), respectively.

C.2 Jacobian of composition of polynomial maps

The goal of this subsection, is to exhibit the structure of the composition of polynomial NN-like maps and their Jacobians. Consider an outer layer of an hPNN, which is denoted as

$$\mathbf{W} \rho_r(q_1, \dots, q_d).$$

In order to see what happens when we substitute variables q_1, \dots, q_d by d_0 -variate polynomials $\mathbf{q}(x_1, \dots, x_{d_0}) \in (\mathcal{H}_{d_0,R})^{\times d}$, we introduce the following definition (which corresponds to (7)):

Definition C.6 (Last layer map). Let $\mathbf{W} \in \mathbb{R}^{n \times d}$ be $n \times d$ matrix r be a number. We define the map ψ that transforms a vector of R -degree d_0 -variate polynomial as follows:

$$\begin{aligned} \psi : (\mathcal{H}_{d_0,R})^{\times d} \times \mathbb{R}^{n \times d} &\rightarrow (\mathcal{H}_{d_0,Rr})^{\times n} \\ (\mathbf{q}(x_1, \dots, x_{d_0}), \mathbf{W}) &\mapsto \psi[\mathbf{q}, \mathbf{W}] := \mathbf{W} \rho_r(\mathbf{q}(x_1, \dots, x_{d_0})), \end{aligned}$$

and denote the Jacobian with respect to the parameters as

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

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

Example C.7. As in Example C.3, we take the case $n = 2$, $d = 2$, $r = 2$, and denote $\mathbf{W} = [\mathbf{w}_1 \quad \mathbf{w}_2]$. Then the last layer map becomes

$$\psi(q_1, q_2) = \mathbf{w}_1 q_1^2 + \mathbf{w}_2 q_2^2.$$

Consider a special case $d_0 = 2$, $R = 2$ so that ψ maps $(q_1, q_2) \in (\mathcal{H}_{2,2})^{\times 2}$ to a vector polynomial in $(\mathcal{H}_{2,4})^{\times 2}$, and let the input polynomials be parameterized as

$$q_j(x_1, x_2) = q_j^{(2,0)} x_1^2 + 2q_j^{(1,1)} x_1 x_2 + q_j^{(0,2)} x_2^2, \quad j \in \{1, 2\},$$

where $q_j^{(i_1, i_2)}$, $(i_1, i_2) \in \{(2, 0), (1, 1), (0, 2)\}$ is the coefficient of q_j next to monomial $x_1^{i_1} x_2^{i_2}$. Then the Jacobian $J_\psi(\mathbf{q}, \mathbf{W})$ is a 10×10 matrix⁸.

⁸since $\dim(\mathcal{H}_{2,4}) = 5$.

The block $J_\psi^{(\mathbf{q})}$ is a 10×6 matrix, whose columns are the 6 polynomials (similarly to (16)):

$$\frac{\partial \psi}{\partial q_j^{(i_1, i_2)}} = 2\mathbf{w}_j x_1^{i_1} x_2^{i_2} (q_j(x_1, x_2)), \quad j \in \{1, 2\}, (i_1, i_2) \in \{(2, 0), (1, 1), (0, 2)\}. \quad (21)$$

In the canonical basis is given as $J_\psi^{(\mathbf{q})} =$

$$(2) \cdot \begin{matrix} \begin{matrix} \mathbf{e}_1 x_1^4 \\ \mathbf{e}_1 x_1^3 x_2 \\ \mathbf{e}_1 x_1^2 x_2^2 \\ \mathbf{e}_1 x_1 x_2^3 \\ \mathbf{e}_1 x_2^4 \\ \mathbf{e}_2 x_1^4 \\ \mathbf{e}_2 x_1^3 x_2 \\ \mathbf{e}_2 x_1^2 x_2^2 \\ \mathbf{e}_2 x_1 x_2^3 \\ \mathbf{e}_2 x_2^4 \end{matrix} & \begin{bmatrix} \frac{\partial \psi}{\partial q_1^{(2,0)}} & \frac{\partial \psi}{\partial q_1^{(1,1)}} & \frac{\partial \psi}{\partial q_1^{(0,2)}} & \frac{\partial \psi}{\partial q_2^{(2,0)}} & \frac{\partial \psi}{\partial q_2^{(1,1)}} & \frac{\partial \psi}{\partial q_2^{(0,2)}} \\ W_{1,1}q_1^{(2,0)} & 0 & 0 & W_{1,2}q_2^{(2,0)} & 0 & 0 \\ W_{1,1}q_1^{(1,1)} & W_{1,1}q_1^{(2,0)} & 0 & W_{1,2}q_2^{(1,1)} & W_{1,2}q_2^{(2,0)} & 0 \\ W_{1,1}q_1^{(0,2)} & W_{1,1}q_1^{(1,1)} & W_{1,1}q_1^{(2,0)} & W_{1,2}q_2^{(0,2)} & W_{1,2}q_2^{(1,1)} & W_{1,2}q_2^{(2,0)} \\ 0 & W_{1,1}q_1^{(0,2)} & W_{1,1}q_1^{(1,1)} & 0 & W_{1,2}q_2^{(0,2)} & W_{1,2}q_2^{(1,1)} \\ 0 & 0 & W_{1,1}q_1^{(0,2)} & 0 & 0 & W_{1,2}q_2^{(0,2)} \\ W_{2,1}q_1^{(2,0)} & 0 & 0 & W_{2,2}q_2^{(2,0)} & 0 & 0 \\ W_{2,1}q_1^{(1,1)} & W_{2,1}q_1^{(2,0)} & 0 & W_{2,2}q_2^{(1,1)} & W_{2,2}q_2^{(2,0)} & 0 \\ W_{2,1}q_1^{(0,2)} & W_{2,1}q_1^{(1,1)} & W_{2,1}q_1^{(2,0)} & W_{2,2}q_2^{(0,2)} & W_{2,2}q_2^{(1,1)} & W_{2,2}q_2^{(2,0)} \\ 0 & W_{2,1}q_1^{(0,2)} & W_{2,1}q_1^{(1,1)} & 0 & W_{2,2}q_2^{(0,2)} & W_{2,2}q_2^{(1,1)} \\ 0 & 0 & W_{2,1}q_1^{(0,2)} & 0 & 0 & W_{2,2}q_2^{(0,2)} \end{bmatrix} \end{bmatrix}$$

The second block, similarly to (17), is a 10×4 matrix whose columns are

$$\frac{\partial \psi}{\partial W_{i,j}} = \mathbf{e}_i (q_j(x_1, x_2))^2, \quad i, j \in \{1, 2\}, \quad (22)$$

and has similar structure to $J_\psi^{(\mathbf{W})}$ in Example C.3.

C.3 A certificate of maximal rank for the Jacobian of the last layer

The following proposition gives a condition for when the Jacobian of the last layer map has maximal rank.

Proposition C.8. *Let $m \leq d_0, d, n$ and $r, R \geq 2$ be fixed, and the matrices $\mathbf{V} \in \mathbb{R}^{d \times m}$ and $\mathbf{W} \in \mathbb{R}^{n \times d}$ be such that the equalities (14)–(15) are satisfied.*

Consider a particular polynomial vector $\hat{\mathbf{q}}(x_1, \dots, x_m) \in (\mathcal{H}_{m,R})^{\times d} \subseteq (\mathcal{H}_{d_0,R})^{\times d}$ defined as

$$\hat{\mathbf{q}}(\mathbf{x}) := \mathbf{V} \begin{bmatrix} x_1^R \\ x_2^R \\ \vdots \\ x_m^R \end{bmatrix}. \quad (23)$$

Then we have that the evaluation of the Jacobian of ψ at the particular point $(\hat{\mathbf{q}}, \mathbf{W})$ is of maximal possible rank, and, in particular,

$$\text{rank}\{J_\psi(\hat{\mathbf{q}}, \mathbf{W})\} = d(n-1) + d \binom{R+d_0-1}{R} \quad (24)$$

and

$$\text{rank}\{J_\psi^{(\mathbf{q})}(\hat{\mathbf{q}}, \mathbf{W})\} = d \binom{R+d_0-1}{R} \quad (25)$$

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

Before proving Proposition C.8, we give an illustrative example of the last layer map.

Example C.9 (Example C.3, continued). *We continue Examples C.3 and C.7. In this case, the vector polynomial $\hat{\mathbf{q}}$ from Proposition C.8 reads*

$$\hat{\mathbf{q}}(x_1, x_2) = \begin{bmatrix} \hat{q}_1(x_1, x_2) \\ \hat{q}_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} (\alpha_1 x_1^2 + \beta_1 x_2^2) \\ (\alpha_2 x_1^2 + \beta_2 x_2^2) \end{bmatrix},$$

i.e., using the notation of Example C.3, the coefficients of the polynomials are

$$\begin{aligned}(\hat{q}_1^{(2,0)}, \hat{q}_1^{(1,1)}, \hat{q}_1^{(0,2)}) &= (\alpha_1, 0, \beta_1), \\(\hat{q}_2^{(2,0)}, \hat{q}_2^{(1,1)}, \hat{q}_2^{(0,2)}) &= (\alpha_2, 0, \beta_2).\end{aligned}$$

Specializing Example C.7 (and removing factor 2 for simlicity), we get

$$\frac{1}{2}J_\psi^{(q)}(\hat{q}, W) = \begin{matrix} & \frac{\partial \psi}{\partial q_1^{(2,0)}} & \frac{\partial \psi}{\partial q_1^{(1,1)}} & \frac{\partial \psi}{\partial q_1^{(0,2)}} & \frac{\partial \psi}{\partial q_2^{(2,0)}} & \frac{\partial \psi}{\partial q_2^{(1,1)}} & \frac{\partial \psi}{\partial q_2^{(0,2)}} \\ \begin{matrix} e_1 x_1^4 \\ e_1 x_1^3 x_2 \\ e_1 x_1^2 x_2^2 \\ e_1 x_1 x_2^3 \\ e_1 x_2^4 \\ e_2 x_1^4 \\ e_2 x_1^3 x_2 \\ e_2 x_1^2 x_2^2 \\ e_2 x_1 x_2^3 \\ e_2 x_2^4 \end{matrix} & \begin{bmatrix} W_{1,1}\alpha_1 & 0 & 0 & W_{1,2}\alpha_2 & 0 & 0 \\ 0 & W_{1,1}\alpha_1 & 0 & 0 & W_{1,2}\alpha_2 & 0 \\ W_{1,1}\beta_1 & 0 & W_{1,1}\alpha_1 & W_{1,2}\beta_2 & 0 & W_{1,2}\alpha_2 \\ 0 & W_{1,1}\beta_1 & 0 & 0 & W_{1,2}\beta_2 & 0 \\ 0 & 0 & W_{1,1}\beta_1 & 0 & 0 & W_{1,2}\beta_2 \\ W_{2,1}\alpha_1 & 0 & 0 & W_{2,2}\alpha_2 & 0 & 0 \\ 0 & W_{2,1}\alpha_1 & 0 & 0 & W_{2,2}\alpha_2 & 0 \\ W_{2,1}\beta_1 & 0 & W_{2,1}\alpha_1 & W_{2,2}\beta_2 & 0 & W_{2,2}\alpha_2 \\ 0 & W_{2,1}\beta_1 & 0 & 0 & W_{2,2}\beta_2 & 0 \\ 0 & 0 & W_{2,1}\beta_1 & 0 & 0 & W_{2,2}\beta_2 \end{bmatrix} \end{matrix}.$$

The matrix $J_\psi^{(W)}$ then, according to (22), becomes

$$J_\psi^{(W)}(\hat{q}, W) = \begin{matrix} & \frac{\partial \psi}{\partial W_{1,1}} & \frac{\partial \psi}{\partial W_{2,1}} & \frac{\partial \psi}{\partial W_{1,2}} & \frac{\partial \psi}{\partial W_{1,2}} \\ \begin{matrix} e_1 x_1^4 \\ e_1 x_1^3 x_2 \\ e_1 x_1^2 x_2^2 \\ e_1 x_1 x_2^3 \\ e_1 x_2^4 \\ e_1 x_1^4 \\ e_1 x_1^3 x_2 \\ e_1 x_1^2 x_2^2 \\ e_1 x_1 x_2^3 \\ e_1 x_2^4 \end{matrix} & \begin{bmatrix} \alpha_1^2 & 0 & \alpha_2^2 & 0 \\ 0 & 0 & 0 & 0 \\ 2\alpha_1\beta_1 & 0 & 2\alpha_2\beta_2 & 0 \\ 0 & 0 & 0 & 0 \\ \beta_1^2 & 0 & \beta_2^2 & 0 \\ 0 & \alpha_1^2 & 0 & \alpha_2^2 \\ 0 & 0 & 0 & 0 \\ 0 & 2\alpha_1\beta_1 & 0 & 2\alpha_2\beta_2 \\ 0 & 0 & 0 & 0 \\ 0 & \beta_1^2 & 0 & \beta_2^2 \end{bmatrix} \end{matrix}.$$

The crux of the proof of Proposition C.8 is the following observation. If we stack together matrices

$\mathbf{J} = \begin{bmatrix} \frac{1}{2}J_\psi^{(q)} & J_\psi^{(W)} \end{bmatrix}$ and permute the columns as follows, we get the block-diagonal matrix

$$\mathbf{J} = \begin{matrix} & \frac{\partial \psi}{\partial q_1^{(2,0)}} & \frac{\partial \psi}{\partial q_1^{(0,2)}} & \frac{\partial \psi}{\partial q_2^{(2,0)}} & \frac{\partial \psi}{\partial q_2^{(0,2)}} & \frac{\partial \psi}{\partial W_{1,1}} & \frac{\partial \psi}{\partial W_{2,1}} & \frac{\partial \psi}{\partial W_{1,2}} & \frac{\partial \psi}{\partial W_{1,2}} & \frac{\partial \psi}{\partial q_1^{(1,1)}} & \frac{\partial \psi}{\partial q_2^{(1,1)}} \\ \begin{matrix} e_1 x_1^4 \\ e_1 x_1^3 x_2 \\ e_1 x_2^4 \\ e_2 x_1^4 \\ e_2 x_1^3 x_2 \\ e_2 x_2^4 \\ e_1 x_1^3 x_2 \\ e_1 x_1 x_2^3 \\ e_2 x_1^3 x_2 \\ e_2 x_1 x_2^3 \end{matrix} & \begin{bmatrix} W_{1,1}\alpha_1 & 0 & W_{1,2}\alpha_2 & 0 & \alpha_1^2 & 0 & \alpha_2^2 & 0 & 0 & 0 \\ W_{1,1}\beta_1 & W_{1,1}\alpha_1 & W_{1,2}\beta_2 & W_{1,2}\alpha_2 & 2\alpha_1\beta_1 & 0 & 2\alpha_1\beta_2 & 0 & 0 & 0 \\ 0 & W_{1,1}\beta_1 & 0 & W_{1,2}\beta_2 & \beta_1^2 & 0 & \beta_2^2 & 0 & 0 & 0 \\ W_{2,1}\alpha_1 & 0 & W_{2,2}\alpha_2 & 0 & 0 & \alpha_1^2 & 0 & \alpha_2^2 & 0 & 0 \\ W_{2,1}\beta_1 & W_{2,1}\alpha_1 & W_{2,2}\beta_2 & W_{2,2}\alpha_2 & 0 & 2\alpha_1\beta_1 & 0 & 2\alpha_1\beta_2 & 0 & 0 \\ 0 & W_{2,1}\beta_1 & 0 & W_{2,2}\beta_2 & 0 & \beta_1^2 & 0 & \beta_2^2 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix}.$$

We see that the top-left block of the matrix \mathbf{J} is nothing but the matrix

$$\begin{bmatrix} \frac{1}{2}J_\varphi^{(V)} & J_\varphi^{(W)} \end{bmatrix},$$

where φ is as in Example C.3, thus it has rank 6. The bottom-right block can be viewed as submatrix $\frac{1}{2}J_\varphi^{(V)}$ (taking first and third columns, for instance), and therefore has full column rank 2.

Thus matrix J_ψ has rank 10 and $J_\psi^{(q)}$ has rank $6 = 4 + 2$.

C.4 Extra notation for the proof of the proposition

In order to prove Proposition C.8 we introduce extra notation for the columns of J_ψ . We first let $\mathbf{W} = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_d]$ as in Remark C.5, so we can express

$$\psi[\mathbf{q}, \mathbf{W}] = \sum_{j=1}^d \mathbf{w}_j(q_j)^r.$$

Already this, similarly to (19) gives us

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

and we denote the linear space spanned by these polynomials (i.e., the range of $J_\psi^{(\mathbf{W})}$) as

$$\mathcal{L}^{(\mathbf{W})} = \text{span} \left\{ \frac{\partial}{\partial W_{i,j}} \psi \right\}_{i,j=1}^{n,d} = \text{range}\{J_\psi^{(\mathbf{W})}\}.$$

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

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

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

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

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

$$\mathbf{f}_{j,\mathbf{i}}(\mathbf{x}) := \frac{\partial \psi}{\partial q_j^{(\mathbf{i})}}(\mathbf{q}, \mathbf{W}) = (r x_1^{i_1} \dots x_{d_0}^{i_{d_0}}) \mathbf{w}_j(q_j)^{r-1}, \quad j = 1, \dots, d, \quad \mathbf{i} \in \mathcal{I}, \quad (26)$$

which are precisely generalizations of (21). We denote the spaces spanned by such polynomials as

$$\mathcal{L}^{(\mathbf{q}, \mathbf{i})} := \text{span}\{\mathbf{f}_{j,\mathbf{i}}(\mathbf{x})\}_{j=1}^d,$$

and their span (the range of $J_\psi^{(\mathbf{q})}$) as

$$\mathcal{L}^{(\mathbf{q})} = \text{span}\{\mathcal{L}^{(\mathbf{q}, \mathbf{i})}\}_{\mathbf{i} \in \mathcal{I}} = \text{range}\{J_\psi^{(\mathbf{q})}\}.$$

Example C.10. In notation of Example C.7, $\mathcal{I} = \{(2, 0), (1, 1), (0, 2)\}$. In this case, we have

$$\begin{aligned} \mathcal{L}^{(\mathbf{q}, (2,0))} &= \text{span} \left\{ \frac{\partial \psi}{\partial q_1^{(2,0)}}, \frac{\partial \psi}{\partial q_2^{(2,0)}} \right\}, \\ \mathcal{L}^{(\mathbf{q}, (1,1))} &= \text{span} \left\{ \frac{\partial \psi}{\partial q_1^{(1,1)}}, \frac{\partial \psi}{\partial q_2^{(1,1)}} \right\}, \\ \mathcal{L}^{(\mathbf{q}, (0,2))} &= \text{span} \left\{ \frac{\partial \psi}{\partial q_1^{(0,2)}}, \frac{\partial \psi}{\partial q_2^{(0,2)}} \right\}, \end{aligned}$$

which correspond to columns $\{1, 4\}$, $\{2, 5\}$, $\{3, 6\}$, respectively, of the matrix $J_\psi^{(\mathbf{q})}$.

Remark C.11. Proving Proposition C.8 (i.e., proving that (24)–(25) hold) is equivalent to showing that

$$\dim \text{span}\{\mathcal{L}^{(\mathbf{q})}, \mathcal{L}^{(\mathbf{W})}\} = d(n-1) + d \binom{R+d_0-1}{R}, \quad (27)$$

$$\dim \mathcal{L}^{(\mathbf{q})} = d \binom{R+d_0-1}{R}, \quad (28)$$

respectively.

The strategy of proving that the dimensions of these subspaces are maximal is to show that the individual subspaces $\mathcal{L}^{(\mathbf{q}, \mathbf{i})}$ are orthogonal under some conditions (which is similar to bringing \mathbf{J} into the block-diagonal form in Example C.9).

C.5 Proof of the proposition: case $m = d_0$

We first prove the proposition for the case when the number of input variables d_0 is equal to the number of variables m used in the certificate.

Proof of Proposition C.8 (case $m = d_0$). Recall that in the notation of the previous subsection we need to calculate

$$\dim \text{span} \left(\mathcal{L}^{(\mathbf{W})}, \text{span} \{ \mathcal{L}^{(\mathbf{q}, i)} \}_{i \in \mathcal{I}} \right).$$

Now let us consider these subspaces for a particular choice of $\mathbf{q} = \hat{\mathbf{q}}$ of the form (23). We have that $\mathbf{f}_{j,i}$ from (26) have the form

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

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

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

with $\mathcal{I}_0 := \{(R, 0, \dots, 0), (0, R, 0, \dots, 0), \dots, (0, 0, \dots, R)\}$. For the same reasons we get

$$\mathcal{L}^{(\mathbf{W})} \perp \mathcal{L}^{(\mathbf{q}, i)} \text{ for all } i \in \mathcal{I} \setminus \mathcal{I}_0.$$

Therefore, we get

$$\text{rank}\{J_\psi\} = \dim \text{span} \left(\mathcal{L}^{(\mathbf{W})}, \text{span} \{ \mathcal{L}^{(\mathbf{q}, i)} \}_{i \in \mathcal{I}_0} \right) + \sum_{i \in \mathcal{I} \setminus \mathcal{I}_0} \dim(\mathcal{L}^{(\mathbf{q}, i)}).$$

Let us look at those dimensions separately. Denote $\mathbf{z} = [z_1 \ \dots \ z_m]^\top$, with

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

so that for $\hat{\mathbf{q}}$ of the form (23) it holds

$$\hat{\mathbf{q}}_j = \mathbf{v}_j^\top \mathbf{z}.$$

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

$$\dim(\mathcal{L}^{(\mathbf{q}, i)}) = \dim \text{span} \left(\{\mathbf{w}_j(\hat{\mathbf{q}}_j)^{r-1}\}_{j=1}^d \right) = d,$$

where the last equality follows from Lemma C.1 and (20).

By doing the same substitution, we obtain that

$$\text{span} \left(\mathcal{L}^{(\mathbf{W})}, \text{span} \{ \mathcal{L}^{(\mathbf{q}, i)} \}_{i \in \mathcal{I}_0} \right) = \text{span} \left(\{e_i(\mathbf{v}_j^\top \mathbf{z})^r\}_{i,j=1}^{n,d}, \{\mathbf{w}_j z_\ell (\mathbf{v}_j^\top \mathbf{z})^{r-1}\}_{j,\ell=1}^{d,m} \right),$$

which is exactly the set of vectors in (19)–(20). Therefore, by Lemma C.1, we have

$$\dim \text{span} \left(\mathcal{L}^{(\mathbf{W})}, \{\mathcal{L}^{(\mathbf{q}, \ell)}\}_{\ell \in \mathcal{I}_0} \right) = (n-1)d + md, \quad \text{and} \quad (29)$$

$$\dim \text{span} \{ \mathcal{L}^{(\mathbf{q}, \ell)} \}_{\ell \in \mathcal{I}_0} = md. \quad (30)$$

Taking into account that

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

this proves (27) for $d_0 = m$. Equality (28) (for $d_0 = m$) can be proved similarly using the fact that

$$\begin{aligned} \text{rank}\{J_\psi^{(\mathbf{q})}(\hat{\mathbf{q}}, \mathbf{W})\} &= \dim \text{span} \{ \mathcal{L}^{(\mathbf{q}, \ell)} \}_{\ell \in \mathcal{I}_0} + \sum_{i \in \mathcal{I} \setminus \mathcal{I}_0} \dim(\mathcal{L}^{(\mathbf{q}, i)}) \\ &= md + d(\#(\mathcal{I}) - \#(\mathcal{I}_0)) = d(\#(\mathcal{I})). \end{aligned}$$

□

C.6 Proof of the proposition: extending to the case of more variables

Proof of Proposition C.8 (case $m < d_0$). We denote by \mathcal{I}_m (with some abuse of notation) the multi-indices that correspond to the monomials that depend only on x_1, \dots, x_m :

$$\mathcal{I}_m = \{i \in \mathcal{I} : i_{m+1}, \dots, i_{d_0} = 0\}$$

and we define

$$\mathcal{L}_m^{(q)} = \text{span}\{\mathcal{L}^{(q,i)}\}_{i \in \mathcal{I}_m}, \quad \mathcal{L}_{ext} = \text{span}\{\mathcal{L}^{(q,i)}\}_{i \in \mathcal{I} \setminus \mathcal{I}_m}.$$

From the first part of the proof (case $m = d_0$), we have already proved that

$$\dim \text{span}(\mathcal{L}_m^{(q)}, \mathcal{L}^{(W)}) = d(n-1) + d \binom{R+m-1}{R}. \quad (31)$$

and

$$\dim \mathcal{L}_m^{(q)} = d \binom{R+m-1}{R}. \quad (32)$$

What is left to show is that adding \mathcal{L}_{ext} to these subspaces does not drop the rank.

Since the particular choice of $q = \hat{q}(x_1, \dots, x_m)$ depends only on the m variables, thanks to (26) we have

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

This immediately implies that $\mathcal{L}^{(q,i)} \perp \mathcal{L}^{(q,\ell)}$ if $(i_{m+1}, \dots, i_{d_0}) \neq (\ell_{m+1}, \dots, \ell_{d_0})$, as well as $\mathcal{L}^{(q,i)} \perp \mathcal{L}^{(W)}$ if $(i_{m+1}, \dots, i_{d_0}) \neq \mathbf{0}$. Therefore, we get

$$\mathcal{L}^{(q)} = \mathcal{L}_m^{(q)} \oplus \mathcal{L}_{ext} \quad \text{and} \quad \text{span}(\mathcal{L}^{(q)}, \mathcal{L}^{(W)}) = \text{span}(\mathcal{L}_m^{(q)}, \mathcal{L}^{(W)}) \oplus \mathcal{L}_{ext},$$

and, consequently, we just need to show that \mathcal{L}_{ext} is of maximal dimension. To show this, we split $\mathcal{I} \setminus \mathcal{I}_m$ into a direct sum according to the degrees of the last $d_0 - m$ variables:

$$\dim \mathcal{L}_{ext} = \sum_{\substack{i_{m+1}, \dots, i_{d_0} \geq 0 \\ 1 \leq i_{m+1} + \dots + i_{d_0} \leq R}} \dim \mathcal{L}^{(q, (*, i_{m+1}, \dots, i_{d_0}))}$$

where

$$\mathcal{L}^{(q, (*, i_{m+1}, \dots, i_{d_0}))} = \text{span}\{\mathcal{L}^{(q, (i_1, \dots, i_m, i_{m+1}, \dots, i_{d_0}))}\}_{(i_1, \dots, i_m) : \substack{i_k \geq 0, \\ i_1 + \dots + i_m = R - i_{m+1} - \dots - i_{d_0}}}.$$

But then, for a fixed $(i_{m+1}, \dots, i_{d_0})$ such $i_{m+1} + \dots + i_{d_0} = R_0 \leq R$, the dimension of this subspace is equal to

$$\begin{aligned} \dim \mathcal{L}^{(q, (*, i_{m+1}, \dots, i_{d_0}))} &= \dim \text{span}\{x_1^{i_1} \dots x_m^{i_m} w_j (\hat{q}_j)^{r-1}\}_{\substack{j=1, \dots, d, \\ i_1, \dots, i_m \geq 0 \\ i_1 + \dots + i_m = R - R_0}} \\ &= \dim \text{span}\{x_1^{i_1} \dots x_m^{i_m} w_j (\hat{q}_j)^{r-1}\}_{\substack{j=1, \dots, d, \\ i_1, \dots, i_m \geq 0 \\ i_1 + \dots + i_m = R - R_0}} \\ &= \dim \text{span}\{x_1^{R_0+i_1} \dots x_m^{i_m} w_j (\hat{q}_j)^{r-1}\}_{\substack{j=1, \dots, d, \\ i_1, \dots, i_m \geq 0 \\ i_1 + \dots + i_m = R - R_0}}, \end{aligned}$$

but the latter set of polynomials is linearly independent because it is a subset of the basis vectors of $\mathcal{L}_m^{(q)}$, which are linearly independent by (32). Therefore we get \mathcal{L}_{ext} is of maximal possible dimension (the spanning columns are linearly independent). \square

C.7 Localization theorem

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

With some abuse of notation, let $\theta = (\mathbf{W}_1, \dots, \mathbf{W}_{L-1})$, so that $\mathbf{w} = (\theta, \mathbf{W}_L)$ and denote $R = r_1 \cdots r_{L-2}$.

Let ψ be as the one defined in Proposition C.8, but given for the last subnetwork, so that $n = d_L, d = d_{L-1}, r = r_{L-1}, \mathbf{W} = \mathbf{W}_L$. Then we have that

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

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

Therefore, by the chain rule

$$\mathbf{J}_{\mathbf{p}_{\mathbf{w}}}(\mathbf{w}) = \left[\underbrace{\left(\mathbf{J}_{\psi}^{(q)} \Big|_{q=h(\theta)} \right)}_{=\mathbf{J}_1(\mathbf{w})} \cdot \mathbf{J}_h(\theta) \quad \underbrace{\mathbf{J}_{\psi}^{(\mathbf{W})} \Big|_{q=h(\theta)}}_{=\mathbf{J}_2(\theta)} \right],$$

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

$$\text{rank}\{\mathbf{J}_h(\theta)\} = \sum_{\ell=0}^{L-2} d_{\ell} d_{\ell+1} - \sum_{\ell=1}^{L-2} d_{\ell}.$$

Now we show the ranks for other matrices. Observe that

$$\text{rank}\left\{ \left[\left(\mathbf{J}_{\psi}^{(q)} \Big|_{q=h(\theta)} \right) \quad \mathbf{J}_{\psi}^{(\mathbf{W})} \Big|_{q=h(\theta)} \right] \right\} \leq d_{L-1} \binom{R + d_0 - 1}{R} + (d_L - 1)d_{L-1} \quad (33)$$

due to the essential ambiguities. But then if we find a particular point $\hat{\theta}$, where rank is maximal for $\hat{q} = h(\hat{\theta})$, then the rank in (33) will be maximal for generic θ .

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

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

and

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

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

$$h(\hat{\theta}) = \mathbf{V} \begin{bmatrix} x_1^R \\ \vdots \\ x_m^R \end{bmatrix},$$

so exactly as in Proposition C.8. Therefore rank in (33) will be maximal for generic (θ, \mathbf{W}_L) and also

$$\text{rank}\left\{ \left(\mathbf{J}_{\psi}^{(q)} \Big|_{q=h(\theta)} \right) \right\} = d_{L-1} \binom{R + d_0 - 1}{R}$$

for generic θ (i.e., the matrix is full rank).

This leads to $\text{rank}\{\mathbf{J}_1(\mathbf{w})\} = \mathbf{J}_h(\boldsymbol{\theta})$ for generic $\boldsymbol{\theta}$. Finally, we have that

$$\begin{aligned} \text{rank}\{\mathbf{J}_{\mathbf{p}_w}(\mathbf{w})\} &= \text{rank}\{\mathbf{J}_1(\boldsymbol{\theta})\} + \text{rank}\{\Pi_{\text{span } \mathbf{J}_1(\boldsymbol{\theta})^\perp} \mathbf{J}_2(\boldsymbol{\theta})\} \\ &\geq \text{rank}\{\mathbf{J}_1(\boldsymbol{\theta})\} + \text{rank}\left\{\Pi_{\text{span}\left(\mathbf{J}_\psi^{(q)}\Big|_{q=h(\boldsymbol{\theta})}\right)^\perp} \mathbf{J}_2(\boldsymbol{\theta})\right\} \\ &= \sum_{\ell=0}^{L-2} d_\ell d_{\ell+1} - \sum_{\ell=1}^{L-2} d_\ell + (d_L - 1)d_{L-1} \\ &= \sum_{\ell=0}^{L-1} d_\ell d_{\ell+1} - \sum_{\ell=1}^{L-1} d_\ell, \end{aligned}$$

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

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

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

C.8 Implications of the localization theorem

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

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

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

\square

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 $\tilde{d}_\ell \geq 2$. But then the Kruskal bound for identifiability of $(\tilde{d}_{\ell-1}, d_\ell, d_{\ell+1})$ is

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

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

Corollary 19 (Identifiability of bottleneck hPNNs) *Consider the “bottleneck” architecture with*

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

and $d_b \geq 2$. Suppose that $r_1, \dots, r_b \geq 2$ and that the decoder part satisfies $\frac{d_\ell}{r_\ell} \leq d_b - 1$ for $\ell \in \{b+1, \dots, L-1\}$. 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 $\tilde{d}_\ell = d_\ell$ and thus identifiability of $(\tilde{d}_{\ell-1}, d_\ell, d_{\ell+1})$ holds for $r_\ell \geq 2$ (the same argument as in the pyramidal case).
- For layers $\ell \in \{b+1, \dots, L\}$ (the decoder part), we have $\tilde{d}_\ell = d_b$ and thus identifiability of $(\tilde{d}_{\ell-1}, d_\ell, d_{\ell+1})$ holds for

$$r_\ell \geq \frac{d_\ell}{d_b - 1},$$

rearranging gives the desired result. \square

D 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 homogeneity. We then prove our main finite identifiability results showing that finite identifiability of 2-layer subnetworks of the homogenized PNNs is sufficient to guarantee the finite identifiability of the original PNN.

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

D.1 The homogeneity procedure: the hPNN associated to a PNN

Our homogeneity 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 $\mathcal{P}_{d,r} \rightarrow \mathcal{H}_{d+1,r}$ by $\text{homog}(\cdot)$, and it acts as follows: for every polynomial $p \in \mathcal{P}_{d,r}$, $\tilde{p} = \text{homog}(p) \in \mathcal{H}_{d+1,r}$ is the unique homogeneous polynomial in $d+1$ variables such that*

$$\tilde{p}(x_1, \dots, x_d, 1) = p(x_1, \dots, x_d).$$

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

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

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

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

which satisfies the required properties. \square

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 $\mathbf{p} \in (\mathcal{P}_{d_0, r_{\text{total}}})^{\times d_L}$ 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 homogeneity $\tilde{\mathbf{p}} = \text{homog}(\mathbf{p})$ admits an hPNN decomposition for the same activation degrees \mathbf{r} and extended $\tilde{\mathbf{d}} = (d_0 + 1, \dots, d_{L-1} + 1, d_L)$, $\tilde{\mathbf{p}} = \text{hPNN}_{\tilde{\mathbf{d}}, \mathbf{r}}[\tilde{\mathbf{w}}]$, $\tilde{\mathbf{w}} = (\tilde{\mathbf{W}}_1, \dots, \tilde{\mathbf{W}}_{L-1}, [\mathbf{W}_L \quad \mathbf{b}_L])$, with matrices $\tilde{\mathbf{W}}_\ell$ for $\ell = 1, \dots, L-1$ given as*

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

Proof of Proposition 23. Denote $\mathbf{p}_1(\mathbf{x}) = \rho_{r_1}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1)$. Let $\tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ z \end{bmatrix} \in \mathbb{R}^{d_0+1}$. Observe first that

$$\rho_{r_1}(\tilde{\mathbf{W}}_1 \tilde{\mathbf{x}}) = \begin{bmatrix} \tilde{\mathbf{p}}_1(\tilde{\mathbf{x}}) \\ z^{r_1} \end{bmatrix}.$$

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

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

$$\tilde{\mathbf{W}}_{2\rho_{r_1}}(\tilde{\mathbf{W}}_1 \tilde{\mathbf{x}}) = \tilde{\mathbf{W}}_2 \begin{bmatrix} \tilde{\mathbf{p}}_1(\tilde{\mathbf{x}}) \\ z^{r_1} \end{bmatrix} = \mathbf{W}_2 \tilde{\mathbf{p}}_1(\tilde{\mathbf{x}}) + z^{r_1} \mathbf{b}_2.$$

Specializing at $z = 1$, we recover

$$\mathbf{W}_2 \mathbf{p}_1(\mathbf{x}) + \mathbf{b}_2 = \mathbf{p}(\mathbf{x}) = \tilde{\mathbf{p}}(\mathbf{x}, 1),$$

hence

$$\tilde{\mathbf{W}}_{2\rho_{r_1}}(\tilde{\mathbf{W}}_1 \tilde{\mathbf{x}}) = \tilde{\mathbf{p}}(\tilde{\mathbf{x}}).$$

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

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

□

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

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

$$\tilde{\mathbf{W}}'_\ell = \begin{cases} \tilde{\mathbf{P}}_\ell \tilde{\mathbf{D}}_\ell \tilde{\mathbf{W}}_\ell \tilde{\mathbf{D}}_{\ell-1}^{-r_{\ell-1}} \tilde{\mathbf{P}}_{\ell-1}^\top, & \ell < L, \\ \tilde{\mathbf{W}}_L \tilde{\mathbf{D}}_{L-1}^{-r_{L-1}} \tilde{\mathbf{P}}_{L-1}^\top, & \ell = L. \end{cases} \quad (34)$$

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

$$\tilde{\mathbf{W}}'_\ell = \begin{cases} \begin{bmatrix} \mathbf{W}'_\ell & \mathbf{b}'_\ell \\ 0 & 1 \end{bmatrix}, & \ell < L, \\ \begin{bmatrix} \mathbf{W}'_L & \mathbf{b}'_L \end{bmatrix}, & \ell = L. \end{cases} \quad (35)$$

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

Next, by Proposition 33, for $\ell = 1, \dots, L-1$ the matrices satisfy $\text{krank}\{(\tilde{\mathbf{W}}'_\ell)^\top\} \geq 2$ (as well as for any equivalent $\text{krank}\{(\tilde{\mathbf{W}}'_\ell)^\top\} \geq 2$). This implies that the matrix $\tilde{\mathbf{W}}'_\ell$ contains only a single row of the form $[0 \cdots 0 \alpha]$ (which is its last row). Therefore in order for $\tilde{\mathbf{W}}'_1$ to be of the form (35), the matrices $\tilde{\mathbf{P}}_1 \tilde{\mathbf{D}}_1$ must be of the form

$$\tilde{\mathbf{P}}_1 = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{\mathbf{D}}_1 = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}.$$

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

$$\tilde{\mathbf{P}}_\ell = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{\mathbf{D}}_\ell = \begin{bmatrix} * & 0 \\ 0 & 1 \end{bmatrix}.$$

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 $\text{hPNN}_r[\tilde{\mathbf{w}}]$ is finite-to-one. Then the same reasoning applies to all alternative (non-equivalent) parameters $\tilde{\mathbf{w}}$ 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. □

D.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 $(\mathbf{W}_2, \mathbf{b}_2, \mathbf{W}_1, \mathbf{b}_1)$, the following hPNN is unique (finite-to-one)

$$p(\tilde{\mathbf{x}}) = [\mathbf{W}_2 \quad \mathbf{b}_2] \rho_{r_1}(\tilde{\mathbf{W}}_1 \tilde{\mathbf{x}}) \quad (36)$$

with $\tilde{\mathbf{W}}_1$ given as

$$\tilde{\mathbf{W}}_1 = \begin{bmatrix} \mathbf{W}_1 & \mathbf{b}_1 \\ 0 & 1 \end{bmatrix}.$$

We see that $\tilde{\mathbf{W}}_1$ 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 $\tilde{\mathbf{W}}_1$ on the right by any nonsingular $(d_0 + 1) \times (d_0 + 1)$ matrix \mathbf{Q} . We note that the image of the polynomial map

$$\mathbb{R}^{(d_0+1) \times (d_0+1)} \times \mathbb{R}^{d_1 \times d_0} \times \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{(d_1+1) \times (d_0+1)} \\ (Q, \mathbf{W}_1, \mathbf{b}_1) \mapsto \tilde{\mathbf{W}}_1 Q,$$

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 $\tilde{\mathbf{W}}_1$ constructed from almost all $(\mathbf{W}_1, \mathbf{b}_1)$ 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 $([\mathbf{W}_2 \quad \mathbf{b}_2], \tilde{\mathbf{W}}_1)$. \square

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

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

Lemma D.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

$$\mathbf{W}_2 = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}, \quad \mathbf{A} \in \mathbb{R}^{n \times d}, \quad \mathbf{B} \in \mathbb{R}^{k \times d}, \quad \mathbf{W}_1,$$

so that

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

But then assume that $p^{(A)}$ is finite-to-one at $(\mathbf{W}_1, \mathbf{A})$ a given parameter. Then by Lemma 31 we have that the elements of $\mathbf{q}_1 = \sigma_r(\mathbf{W}_1 \mathbf{x})$ are linearly independent, hence \mathbf{B} has a unique solution from the linear system

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

Note that for $(\mathbf{W}_1, \mathbf{W}_2)$, the subset of parameters $(\mathbf{W}_1, \mathbf{A})$ is also generic, hence global (resp. finite) identifiability for widths (m, d, n) implies global (resp. finite) identifiability for widths $(m, d, n+k)$. \square

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

$$(d_0 + 1, \dots, d_{L-1} + 1, d_L) \quad \text{and} \quad (d_0 + 1, \dots, d_{L-1} + 1, d_L + 1)$$

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 D.1 for the architecture $(d_{\ell-1} + 1, d_\ell + 1, d_{\ell+1} + 1)$.
- **Induction step: $(L = k - 1) \rightarrow (L = k)$** Assume that the statement holds for $L = k - 1$. Now consider the case $L = k$. As in the proof of Theorem 11, we set $\tilde{\theta} = (\tilde{\mathbf{W}}_1, \dots, \tilde{\mathbf{W}}_{L-1})$, so that $\tilde{\mathbf{w}} = (\tilde{\theta}, \tilde{\mathbf{W}}_L)$ and denote $R = r_1 \cdots r_{L-2}$. The difference is that the parameters are now $\tilde{\theta} := \tilde{\theta}(\theta)$, where

$$\theta = (\mathbf{W}_1, \dots, \mathbf{W}_{L-1}, \mathbf{b}_1, \dots, \mathbf{b}_{L-1}).$$

Let ψ be as the one defined in Proposition C.8, but given for the last subnetwork, so that $n = d_L, d = d_{L-1} + 1, r = r_{L-1}, \mathbf{W} = \tilde{\mathbf{W}}_L$. Then we have that

$$p_{\tilde{\mathbf{w}}}(\tilde{\theta}(\theta), \mathbf{W}) = \psi[h(\tilde{\theta}(\theta)), \mathbf{W}_L],$$

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

Again, by the chain rule

$$\mathbf{J}_{p_{\tilde{\mathbf{w}}}}(\mathbf{w}) = \left[\underbrace{\left(\mathbf{J}_{\psi}^{(q)} \Big|_{q=h(\tilde{\theta}(\theta))} \right)}_{=\mathbf{J}_1(\tilde{\mathbf{w}})} \cdot \mathbf{J}_h(\tilde{\theta}(\theta)) \quad \underbrace{\mathbf{J}_{\psi}^{(\mathbf{W})} \Big|_{q=h(\tilde{\theta}(\theta))}}_{=\mathbf{J}_2(\theta)} \right] \begin{bmatrix} \mathbf{J}_{\tilde{\theta}}(\theta) & \mathbf{I} \end{bmatrix},$$

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 $\tilde{\theta} = \tilde{\theta}(\theta)$. But for this remark that we can use almost the same construction example Proposition C.8, but choosing slightly different matrices: $\tilde{\theta}' = (\widehat{\mathbf{W}}'_1, \dots, \widehat{\mathbf{W}}'_{L-1})$ with

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

and

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

where in Lemma C.1 we can choose generic \mathbf{V}' structured as

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

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

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

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

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

then the L -layer PNN with architecture (\mathbf{d}, \mathbf{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. □

D.3 Truncation of PNNs with bias terms

In this appendix, we describe an alternative (to homogenization) approach to prove the identifiability of the weights \mathbf{W}_ℓ of $\text{PNN}_{\mathbf{d},\mathbf{r}}[(\mathbf{w}, \mathbf{b})]$ 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 \mathbf{b}_ℓ .

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

Example D.2. 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 $q = \text{lt}\{p\} \in \mathcal{H}_{2,2}$ becomes

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

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

Lemma D.3. Let $\mathbf{p} = \text{PNN}_{\mathbf{d},\mathbf{r}}[(\mathbf{w}, \mathbf{b})]$ be a PNN with bias terms. Then its truncation is the hPNN with the same weight matrices

$$\text{lt}\{\mathbf{p}\} = \text{hPNN}_{\mathbf{d},\mathbf{r}}[\mathbf{w}].$$

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

Example D.4. Consider a 2-layer PNN

$$f(\mathbf{x}) = \mathbf{W}_2 \rho_{r_1}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2. \quad (37)$$

Then its truncation is given by

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

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 D.5. Thanks to Lemma D.3, 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.

E Localization theorem: necessary and sufficient conditions for identifiability

This appendix has been added to the camera ready version on the request of the program committee. It explains the changes between the original submission and the camera-ready version.

Our main technical result in Theorem 11 gives sufficient conditions for finite identifiability of deep hPNNs based on identifiability of two-layer subnetworks. In an earlier (submitted) version of the paper, the following results were claimed.

Claim (A, specific uniqueness). Let $\text{hPNN}_{\mathbf{r}}[\mathbf{w}]$, $\mathbf{w} = (\mathbf{W}_1, \dots, \mathbf{W}_L)$ be an L -layer hPNN with architecture (\mathbf{d}, \mathbf{r}) satisfying $d_0, \dots, d_{L-1} \geq 2$, $d_L \geq 1$ and $r_1, \dots, r_L \geq 2$. Then, $\text{hPNN}_{\mathbf{r}}[\mathbf{w}]$ is unique according to Definition 6 if and only if for every $\ell = 1, \dots, L-1$ the 2-layer subnetwork $\text{hPNN}_{(r_\ell)}[(\mathbf{W}_\ell, \mathbf{W}_{\ell+1})]$ is unique as well.

This strong claim implied another claim on identifiability of hPNN architectures, which can be seen as a counterpart of the current Theorem 11.

Claim (B, identifiability). The L -layer hPNN with architecture (\mathbf{d}, \mathbf{r}) satisfying $d_0, \dots, d_{L-1} \geq 2$, $d_L \geq 1$ and $r_1, \dots, r_L \geq 2$ is identifiable according to Definition 8 if and only if for every $\ell = 1, \dots, L-1$ the 2-layer subnetwork with architecture $((d_{\ell-1}, d_\ell, d_{\ell+1}), (r_\ell))$ is identifiable as well.

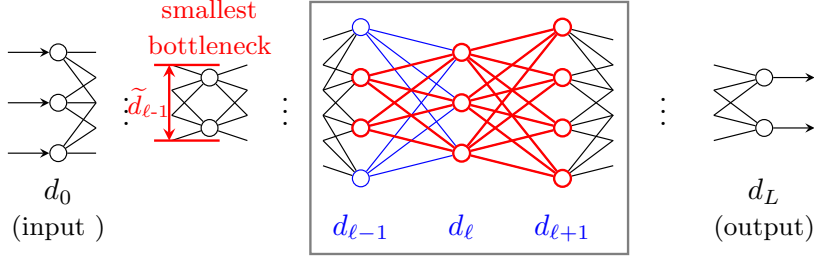


Figure 1: (from NeurIPS poster) Necessary and sufficient conditions for identifiability of an L -layer PNN. **Blue**: necessary conditions, i.e., “only if” part of claim B (identifiability of the $((d_{\ell-1}, d_{\ell}, d_{\ell+1}), (r_{\ell}))$ subnetwork). **Red**: sufficient condition as given by Theorem 11 (identifiability of the $((d_{\ell-1}, d_{\ell}, d_{\ell+1}), (r_{\ell}))$ subnetwork).

We note that the “only if” part always holds, as non-uniqueness of any 2-layer subnetwork implies non-uniqueness of the overall networks. The relation between necessary and sufficient conditions for identifiability is illustrated in Fig. 1.

Thus, both claims (A) and (B) were in fact aiming to answer the following questions:

- (A) Does uniqueness of all 2-layer subnetworks $\text{hPNN}_{(r_{\ell})}[(\mathbf{W}_{\ell}, \mathbf{W}_{\ell+1})]$ imply uniqueness of the overall network $\text{hPNN}_{(r_1, \dots, r_{L-1})}[(\mathbf{W}_1, \dots, \mathbf{W}_L)]$?
- (B) Does identifiability of all $((d_{\ell-1}, d_{\ell}, d_{\ell+1}), (r_{\ell}))$ 2-layer architectures imply the identifiability of the overall architecture $((d_0, \dots, d_L), (r_1, \dots, r_{L-1}))$?

We show below that the answer to this question is negative, both for specific uniqueness (uniqueness of a particular choice of parameters) and generic uniqueness (identifiability of a given architecture), which motivated the update of the paper.

E.1 Supporting examples

Absence of specific uniqueness (counterexample to claim (A)). Consider the simplest architecture with $\mathbf{d} = (2, 2, 2)$, $\mathbf{r} = (2, 2)$, for which the conditions of Theorem 11 are verified due to Proposition 12. Example 41 from the last section of the paper provides an example of specific network of the format (\mathbf{d}, \mathbf{r}) violating claim (A). We provide below an expanded version of this example.

Example 41 (No specific uniqueness). Consider two polynomials:

$$\mathbf{p}(x_1, x_2) = \begin{bmatrix} (x_1^2 + x_2^2)^2 \\ (x_1^2 - x_2^2)^2 \end{bmatrix}.$$

Note that $\begin{bmatrix} x_1^2 & x_2^2 \end{bmatrix}^T = \rho_2(x_1, x_2)$, therefore this polynomial vector can be written as

$$\mathbf{p}(\mathbf{x}) = \rho_2(\mathbf{W}_2 \rho_2(\mathbf{x}))$$

for the following choice of weight matrix:

$$\mathbf{W}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

so that we have

$$\begin{aligned} \mathbf{p}(\mathbf{x}) &= \mathbf{I}_2 \rho_2(\mathbf{W}_2 \mathbf{I}_2 \rho_2(\mathbf{x})) \\ &= \text{hPNN}_{(2,2)}[(\mathbf{I}_2, \mathbf{W}_2, \mathbf{I}_2)], \end{aligned}$$

where \mathbf{I}_2 is the identity matrix. On the other hand, we can use the expansions

$$\begin{aligned} x_1^2 + x_2^2 &= \frac{(x_1 + x_2)^2 + (x_1 - x_2)^2}{2} \\ 2x_1x_2 &= \frac{(x_1 + x_2)^2 - (x_1 - x_2)^2}{2} \end{aligned}$$

and the fact that

$$(x_1^2 - x_2^2) = (x_1^2 + x_2^2)^2 - (2x_1x_2)^2$$

to show that there exists an alternative hPNN expansion of $p(\mathbf{x})$, summarized as

$$p(\mathbf{x}) = \mathbf{W}_3 \rho_2 \left(\frac{1}{2} \mathbf{W}_2 \rho_2(\mathbf{W}_2 \mathbf{x}) \right) = \text{hPNN}_{(2,2)}[(\mathbf{W}_2, \frac{1}{2} \mathbf{W}_2, \mathbf{W}_3)],$$

where

$$\mathbf{W}_3 = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}.$$

We see that the two representations are not equivalent: $(\mathbf{I}_2, \mathbf{W}_2, \mathbf{I}_2) \not\sim (\mathbf{W}_2, \frac{1}{2} \mathbf{W}_2, \mathbf{W}_3)$, as \mathbf{W}_2 cannot be obtained from scaling and permutations of rows of \mathbf{I}_2 .

On the other hand, all the matrices in the expansions $(\mathbf{I}_2, \mathbf{W}_2, \mathbf{W}_3)$ are 2×2 invertible and thus, for example, the networks $\text{hPNN}_{(2)}[(\mathbf{I}_2, \mathbf{W}_2)]$ and $\text{hPNN}_{(2)}[(\mathbf{W}_2, \mathbf{I}_2)]$ have unique representations (similarly to Example 7). More precisely, all the matrices $(\mathbf{I}_2, \mathbf{W}_2, \mathbf{W}_3)$ as well as their transposes have their rank and Kruskal rank both equal to 2, and therefore the conditions of Lemma B.1 are satisfied.

Absence of generic uniqueness (counterexample to claim (B)). Example 41 are not just an isolated example that can be circumvented by looking at a generic parameter set, as shown in the following example.

Example E.1 (No generic identifiability without further assumptions). *We provide a counter example to the conjecture that localization holds in full generality in the generic sense based on the count of dimension. Consider the following architecture:*

$$\mathbf{d} = (2, 3, 3, 1) \quad \text{and} \quad \mathbf{r} = (3, 3).$$

It is easy to see that the subnetworks $((d_0, d_1, d_2), r_1)$ and $((d_1, d_2, d_3), r_2)$ both satisfy the Kruskal-based criterion in Proposition 12 as

$$3 \geq \frac{2d_1 - \min(d_2, d_1)}{\min(d_1, d_0) - 1} = 3, \quad 3 \geq \frac{2d_2 - \min(d_3, d_2)}{\min(d_2, d_1) - 1} = \frac{5}{2},$$

so both subnetworks are identifiable. However, due to Proposition 10, for the global network (\mathbf{d}, \mathbf{r}) to be identifiable the dimension of its associated neurovariety must be equal to

$$d_0 d_1 + d_1 d_2 + d_2 d_3 - d_1 - d_2 = 12.$$

However, the image of $\text{hPNN}_{(\mathbf{d}, \mathbf{r})}[\cdot]$ is in the space degree-9 homogeneous bivariate polynomials, and therefore the neuromanifold (and the neurovariety) lie in $\mathcal{H}_{2,9}$. But $\mathcal{H}_{2,9}$ has dimension 10, thus we arrive at a contradiction with the identifiability of the 3-layer network.

E.2 Statement of changes

In the camera-ready version, the claims (A) and (B) have been replaced with Theorem 11 which uses a stricter condition. This replacement preserves the main conclusions and contributions of the original paper, notably:

1. *The localization of identifiability:* identifiability of 2-layer subnetworks (composed by two consecutive layers) is sufficient to guarantee identifiability of a deep L -layer polynomial network;
2. As a consequence, *uniqueness theorems for tensors can be leveraged* to prove identifiability of deep PNNs; for example, well-known Kruskal theorems imply:
 - a) that pyramidal networks (and their generalizations) are identifiable in degrees ≥ 2 ;
 - b) linear bounds on the so-called *activation degree thresholds* (i.e., identifiability holds for degrees linear in the layer widths);
3. *Identifiability of networks with biases is implied by identifiability of (augmented) bias-free PNN architectures.*

Drawbacks: Despite the fact that our main conclusions still hold, the amended version of the localization theorem lead to the following changes:

- The theorem and the corresponding corollaries for deep architectures concern generic properties (and not specific) and finite identifiability (instead of global identifiability).
- Theorem 11 requires a stronger assumption on 2-layer subnetworks: not only each 2-layer block needs to be identifiable, but also with a possibly smaller number of inputs.
- This stronger condition weakened the result for networks with a bottleneck layer, but keeps the same conclusion (that is, a decoder network needs to have higher degrees compared to the encoder in order to allow for increasing the layer widths).

In the following, we explain the mistake in the original proof of Theorem 11 and discuss the current challenges to extending the amended proof to the localization of global identifiability.

E.3 Remark on the mistake in the original proof and related problems

The mistake in the original proof of Theorem 11 concerned equations (11)–(13) in the original paper (Section A.2.2 of the original supplementary materials), in the induction step of the theorem (going from $L - 1$ to L layers). The original argument is based on constructing the polynomial vector $p'_w(z)$ using the flattening operation $x^{\otimes r'} \mapsto z, \mathcal{H}_{d_0, r'} \rightarrow \mathbb{R}^{(d_0)r'}$. The issue is that the equation (12, original paper) is only valid on a subset of z (z structured as a tensor power) and thus does not imply (13, original paper) as we originally claimed. The absence of this implication broke the inductive argument, requiring the proof to be amended.

An interpretation of this issue is that the flattening destroys the structure from lower layers. In fact, the flattening mapping corresponds to a projection appearing in the computation of the decomposition of polynomials as sums of powers of forms (see the commutative diagram in [92, Section 4]), which makes such a computation (decomposition as a sum of powers of polynomials) very difficult and currently an open problem in general, unless additional knowledge can be used [93].

Our new proof still proceeds similarly by induction (going from $L - 1$ to L layers), where the induction step is related to showing non-defectivity (finite identifiability) of a subvariety of variety of powers of forms, thus connecting to subtle questions in algebraic geometry, such as Fröberg’s conjecture [92] (the latter not solved in full generality, see [85] for an account of recent progress). Extending finite identifiability to identifiability seems challenging, at least with the techniques we are aware of; very recent work in algebraic geometry [75, 86] shows that this transition (i.e., *finite identifiability implies global identifiability*) is possible for the so-called X-rank decompositions, but this result is only applicable to shallow polynomial networks. We are not aware of any systematic progress in the direction of non-additive structures, of which deep PNNs is a special case. Thus, the transition from finite to global identifiability of deep PNNs was left as an open conjecture (Conjecture 40) in the camera-ready version of the paper. We hope that future progress in the field of algebraic geometry will provide the adequate tools to settle this challenging problem⁹.

⁹While preparing the update of the camera-ready paper, we became aware of a recent preprint [94] that claims to prove a much stronger (in many cases) result than Theorem 11 and claims global identifiability as well.