An Embedding of ReLU Networks and an Analysis of their Identifiability

Pierre Stock, Rémi Gribonval

AN EMBEDDING OF RELU NETWORKS AND AN ANALYSIS OF THEIR IDENTIFIABILITY

PIERRE STOCK AND RÉMI GRIBONVAL

Abstract. Neural networks with the Rectified Linear Unit (ReLU) nonlinearity are described by a vector of parameters θ, and realized as a piecewise linear continuous function $\boldsymbol{R}_\theta : x \in \mathbb{R}^d \mapsto \boldsymbol{R}_\theta(x) \in \mathbb{R}^k$. Natural scalings and permutations operations on the parameters θ leave the realization unchanged, leading to equivalence classes of parameters that yield the same realization. These considerations in turn lead to the notion of identifiability – the ability to recover (the equivalence class of) θ from the sole knowledge of its realization $\boldsymbol{R}_\theta$. The overall objective of this paper is to introduce an embedding for ReLU neural networks of any depth, $\boldsymbol{\Phi}(\theta)$, that is invariant to scalings and that provides a locally linear parameterization of the realization of the network. Leveraging these two key properties, we derive some conditions under which a deep ReLU network is indeed locally identifiable from the knowledge of the realization on a finite set of samples $x_i \in \mathbb{R}^d$. We study the shallow case in more depth, establishing necessary and sufficient conditions for the network to be identifiable from a bounded subset $\mathcal{X} \subseteq \mathbb{R}^d$.

Contents

1. Introduction

The empirical success of Deep Neural Networks (DNNs) for traditional machine learning tasks such as image classification is a well-known fact for the research community [1]. While this empirical success percolates to areas ranging from protein folding to symbolic mathematics, a second well-known fact is that the theoretical tools to grasp DNNs and uncover the reasons of their success are still lagging behind the fast-paced experimental results. We argue that a deeper understanding of the expressivity and stability properties of such networks could lead to practical improvements [2, 3]. In this paper, we introduce an embedding, $\mathbf{\Phi}(\theta)$, of the vector θ of network parameters (weights and biases) that exhibits interesting properties for networks based on the popular Rectified Linear Unit (ReLU): in particular, $\mathbf{\Phi}(\theta)$ is invariant to natural rescalings of the parameters that leave unchanged the function implemented by the network. To showcase the potential of this tool, we leverage it to study the expressivity of DNNs from the perspective of their functional equivalence classes.

In the remainder of this section, we first list the papers tackling identifiability of neural networks with a given non-linearity function. We next present some related work that construct an embedding for ReLU networks. Finally, we list applications directly or indirectly derived from the two previous theoretical considerations.

Around Functional Identifiability. First, various results dating back from the 90's identify conditions that allow to identify neural networks *with one hidden layer* equipped with various non-linearities[1] like the hyperbolic tangent [4, 5, 6, 7]. Such results *do not* encompass the ReLU case. Simultaneously, Fefferman derived identifiability conditions for deep networks equipped with the tanh nonlinearity using complex analysis [8]. More recently, the work of Rolnick and Kording [9] reflects a renewed interest for this subject and its application to ReLU networks. The authors propose to reverse-engineer deep ReLU networks and present a constructive algorithm that samples network realizations $\boldsymbol{R}_\theta(x)$ for carefully chosen input points x to deduce the architecture of the network and its parameters, up to rescalings and permutations. The authors prove that their algorithm terminates, except for a measure-zero set of parameters[2]. Similarly, Fornasier *et al.* [10] propose to recover the parameters of a *two-hidden-layer* neural network with smooth nonlinearity by actively sampling finite difference approximations to Hessians of the network, and by combining the insights gained from the sampling with a heuristic for precise attribution of the parameters to the architecture. The authors demonstrate the empirical effectiveness of their approach and claim that the proposed method can be generalized to networks with any depth. Finally, Phuong and Lampert provide a result related to identifiability for ReLU networks under some more restrictive assumptions [11].

Embeddings Zoology. We list here the neural network embeddings in the literature that are the closest to our own embedding, $\boldsymbol{\Phi}(\theta)$, which is introduced in Definition 6. Its main property is that it is invariant under the action of rescalings, as stated more formally in Theorem 1. Schematically, $\boldsymbol{\Phi}(\theta)$ lives in the linear space indexed by network *paths* and each coordinate is a product of weights and/or biases along a particular network path. In a similar fashion, Malgouyres and Landsberg [12] consider a particular class of *linear* structured networks called *Deep Structured Neural Networks*, without biases, and consider only layer-wise rescalings. In [12, Section 6], the authors provide sufficient and necessary conditions for local identifiability by studying complex algebraic varieties leveraging the *Segre embedding* of such networks. The Segre embedding bears a resemblance with $\boldsymbol{\Phi}(\theta)$ since it is also made of product of network parameters, but it does not encompass the biases. Moreover, we consider neuron-wise rescalings in this paper as opposed to less general layer-wise rescalings considered by the authors, and also emcompass the ReLU non-linearity in our approach. Malgouyres later leverages the Segre embedding to study local stability properties of linear neural networks [13]. Finally, Neyshabur *et al.* introduce a family of *path regularizers* to derive an optimization procedure that takes the invariance of the realization $\boldsymbol{R}_\theta$ under the action of the rescalings into account and called Path-SGD [14, 15]. Such path regularizers are scalars – as opposed to vectorial embeddings – that are obtained by summing, for any network path, the norm of the product of all weights along this paths. Moreover, this approach does not take the biases into account, as opposed to our embedding, $\boldsymbol{\Phi}(\theta)$.

[1]It should be noted that the functional equivalence class generally depends on the considered non-linearity. For instance, with the hyperbolic tangent, the authors only consider permutations and sign flips.

[2]This measure-zero set of parameters is not explicitly described by the authors.

Applications. As illustrated with Path-SGD [15], several papers attempt to perform the optimization in the space of network parameters *quotiented* by the rescaling operation. Several work follow and perform the optimization by alternating between a standard SGD step and a projection step that modifies the rescaling coefficients without changing the function implemented by the network [16, 17, 18, 19]. The main difference between these papers is the projection step, that is performed either implicitly (with a regularizer) or explicitly (by computing the optimal[3] rescaling coefficients). In the latter case, the proposed empirical methods may not yield the optimal rescaling coefficients but rather more or less stable and good approximations. Another advantage of rescalings is to improve post-training scalar quantization of neural networks by carefully selecting the rescaling coefficients such that the dynamical range of the weights wighin a layer is relatively small, with as few outliers as possible [20, 21]. More related to the concept of (local) identifiability, Carlini *et al.* [22] design a differential attack to efficiently recover the parameters of remote model up to floating point precision, by sending carefully designed queries x to the remote network and receiving only its output.

After introducing the main notations, we define $\mathbf{\Phi}(\theta)$ and state the main results of the paper in Section 2. Then, we formally state and prove the main properties of the embedding $\mathbf{\Phi}(\theta)$ in Section 3. In particular, we prove that $\mathbf{\Phi}(\theta)$ is invariant under the action of the rescalings. Next, we leverage this embedding to derive partial and local identifiability results for ReLU neural networks of any depth in Section 4. To further demonstrate the validity of our approach, we fully study the shallow case in Section 5 and provide conditions under which a ReLU neural network with one hidden layer is identifiable. Finally, we argue that $\mathbf{\Phi}(\theta)$ may be leveraged to tackle other open problems in the Machine Learning community.

2. General setting and main results

We consider fully-connected feedforward ReLU neural networks with $L \geq 2$ affine layers. Each network is supported on a graph $G = (E, V)$ with vertex set V composed of neurons ν and edge set E composed of connections. The set of neurons V is partitioned into the input layer N_0, $L-1$ hidden layers N_ℓ, $1 \leq \ell \leq L-1$, and the output layer N_L. Hidden neurons compose the set $H = \cup_{\ell=1}^{L-1} N_\ell$. Since we focus on fully-connected networks, the set of connections E is made of all oriented edges $e = \nu \to \nu'$ between neurons belonging to consecutive layers, $\nu \in N_{\ell-1}$, $\nu' \in N_\ell$ for some $1 \leq \ell \leq L$. The subset of incoming edges of neuron ν is denoted $\bullet \to \nu$, while $\nu \to \bullet$ denotes its set of outgoing edges.

Each edge $e \in E$ is equipped with a weight w_e and each hidden neuron $\nu \in H$ with a bias b_ν. Output neurons, i.e. neurons from the last layer $\eta \in N_L$, are also equipped with a bias b_η, which is sometimes constrained to be zero. The set of all neurons equipped with biases is $\bar{H} := H \cup N_L$. Parameters (weights and biases) are gathered in a parameter vector $\theta \in \mathbb{R}^{E \cup \bar{H}}$ where $E \cup \bar{H}$ indexes all possible weights and biases including biases on the output layer. For brevity we may denote θ_e for weights and θ_ν for biases. When

[3] In the sense that such coefficients globally minimize a given objective function.

needed we also write $\theta = (\theta_i)_{i \in E \cup \bar{H}}$. Since we consider a fully connected architecture (this does not prevent some weights to possibly vanish on some edges), θ can also be represented as a set of L matrices $\boldsymbol{W}_\ell = (w_{\nu \to \nu'})_{\nu' \in N_\ell, \nu \in N_{\ell-1}} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, $1 \leq \ell \leq L$ and L vectors $\boldsymbol{b}_\ell = (b_\nu)_{\nu \in N_\ell} \in \mathbb{R}^{N_\ell}$, $1 \leq \ell \leq L$.

2.1. **Network architectures.** Many of the notions of parameter identifiability or non-degeneracy that will be considered are relative to a choice of network "architecture". This is represented both by the graph G (which determines how many layers there are, and how wide they are) but also by a possibly restricted set $\Theta \subseteq \mathbb{R}^{E \cup \bar{H}}$ of network parameters, which may for example account for the following type of constraints:

- restricting to a convolutional structure;
- restricting to sparse networks, possibly with structured sparsity patterns;
- restricting to networks without output biases ($b_\eta = 0$ for every $\eta \in N_L$);
- restricting to networks without biases ($b_\nu = 0$ for every $\nu \in H \cup N_L$).

2.2. **Realization of a network.** Given a parameter θ and an input vector $x \in \mathbb{R}^{N_0}$, we sequentially define $\boldsymbol{y}_0(\theta, x) = x$ and for each $1 \leq \ell \leq L-1$ the pre-activation $\boldsymbol{z}_\ell(\theta, x) = \boldsymbol{W}_\ell \boldsymbol{y}_{\ell-1}(\theta, x) + \boldsymbol{b}_\ell \in \mathbb{R}^{N_\ell}$, the post-activation $\boldsymbol{y}_\ell(\theta, x) = \mathtt{ReLU}(\boldsymbol{z}_\ell(\theta, x)) \in \mathbb{R}^{N_\ell}$ where the rectified linear unit (ReLU) activation function, $\mathtt{ReLU}(t) = \max(t, 0)$, is applied entrywise. Finally we define the realization of the network as the function $\boldsymbol{R}_\theta : x \mapsto \boldsymbol{R}_\theta(x) := \boldsymbol{z}_L(\theta, x) = \boldsymbol{W}_L \boldsymbol{y}_{L-1}(\theta, x) + \boldsymbol{b}_L \in \mathbb{R}^{N_L}$. When needed we will use neuronwise versions of these notations, e.g. $y_\nu(\theta, x) = (\boldsymbol{y}_\ell(\theta, x))_\nu$ where $\nu \in N_\ell$. Note the general convention to denote scalar-valued quantities in plain font to distinguish them from quantities that can be vector-valued, which are generally denoted in bold.

2.3. **Invariance to permutation and scaling.** A well known fact [23] is that the realization of any ReLU-network is invariant to permutations and scalings of the parameter θ. The invariance to permutations is not specific to ReLU-networks, while the scaling-invariance is due to the homogeneity of the ReLU: $\mathtt{ReLU}(\lambda \cdot) = \lambda \mathtt{ReLU}(\cdot)$ for every $\lambda > 0$ and is also valid for other variants such as the leaky-ReLU. While various definitions co-exist in the literature [24, 25, 26], it is convenient to focus first on the practical *per-neuron* rescaling equivalence [23] as stated below.

Rescaling equivalence. Let $\nu \in H$ and $\lambda_\nu > 0$. A neuron-wise scaling multiplies the incoming weights and the bias of ν by λ_ν, and divides the outgoing weights by λ_ν. It is formally defined as $s_{\nu,\lambda_\nu} : \theta = (w, b) \mapsto \theta' = (w', b')$ where for every connection $e \in E$,

$$\forall e \in E, \quad w'_e = \begin{cases} w_e \lambda_\nu & \text{if } e \in \bullet \to \nu \\ \frac{1}{\lambda_\nu} w_e & \text{if } e \in \nu \to \bullet \\ w_e & \text{otherwise,} \end{cases} \qquad \forall \nu \in H, \quad b'_\nu = b_\nu \lambda_\nu. \tag{1}$$

Let $\mathcal{S}$ be the set of neuron-wise scalings. We observe that neuron-wise rescalings commute and are invertible, the inverse of s_{ν,λ_ν} being $s_{\nu,1/\lambda_\nu}$. Let $\langle \mathcal{S} \rangle$ be the commutative group

generated by $\mathcal{S}$. Every $s \in \langle \mathcal{S} \rangle$ can be uniquely represented as the composition

$$s = \bigcirc_{\nu \in H} s_{\nu,\lambda_\nu}$$

where the λ_ν are strictly positive. Note that in this representation, every hidden neuron ν is associated to exactly one neuron-wise rescaling λ_ν.

Definition 1. *θ and θ' are rescaling equivalent if there exists $s \in \langle \mathcal{S} \rangle$ such that $\theta' = s(\theta)$. We then denote $\theta \sim_S \theta'$.*

Notice that if $\theta' \sim_S \theta$, then the output biases are equal: $\theta'_\eta = \theta_\eta$ for all $\eta \in N_L$.

Fact 1. *$\theta' \sim_S \theta$ if, and only if, there exists diagonal matrices $\boldsymbol{\Lambda}_\ell \in \mathbb{R}^{N_\ell \times N_\ell}$ with positive entries, $0 \leq \ell \leq L$ such that $\boldsymbol{\Lambda}_0 = \boldsymbol{I}_{N_0}$, $\boldsymbol{\Lambda}_L = \boldsymbol{I}_{N_L}$, and for every layer $1 \leq \ell \leq L$*

$$\boldsymbol{W}'_\ell = \boldsymbol{\Lambda}_\ell \boldsymbol{W}_\ell \boldsymbol{\Lambda}_{\ell-1}^{-1} \text{ and } \boldsymbol{b}'_\ell = \boldsymbol{\Lambda}_\ell \boldsymbol{b}_\ell. \tag{2}$$

Permutation equivalence. Consider $\pi := (\pi_1, \ldots, \pi_\ell)$ where $\pi_\ell \in \mathfrak{S}_{N_\ell}$ is a permutation of the ℓ-th hidden layer (input and output layers are never permuted), $1 \leq \ell \leq L-1$. Denote $\mathfrak{S}_G = \mathfrak{S}_{N_1} \times \ldots \times \mathfrak{S}_{N_{L-1}}$ the group of all such tuples of permutations. One can define a natural action of the group $\mathfrak{S}_G$ on parameterizations via $\theta \mapsto \pi \circ \theta := \theta'$ where each weight matrix $\boldsymbol{W}'_\ell$ is obtained from $\boldsymbol{W}_\ell$ by permuting rows according to π_ℓ and columns according to $\pi_{\ell-1}$, while bias vector $\boldsymbol{b}'_\ell$ is a permuted version of $\boldsymbol{b}_\ell$ according to π_ℓ.

Definition 2. *Two parameters θ, θ' are* permutation-equivalent *if, and only if, there exists $\pi \in \mathfrak{S}_G$ such that $\theta' = \pi \circ \theta$. This is denoted $\theta \sim_P \theta'$.*
The parameters are permutation-scaling equivalent *if, and only if, there exists θ'' such that $\theta \sim_S \theta'' \sim_P \theta'$. This is denoted $\theta \sim_{PS} \theta'$.*
The parameters are scaling-permutation equivalent *if, and only if, there exists θ'' such that $\theta \sim_P \theta'' \sim_S \theta'$. This is denoted $\theta \sim_{SP} \theta'$.*

Fact 2. *$\theta' \sim_{PS} \theta$ if, and only if, $\theta' \sim_{SP} \theta$, if and only if there exists diagonal matrices $\boldsymbol{\Lambda}_\ell \in \mathbb{R}^{N_\ell \times N_\ell}$ with positive entries and permutation matrices $\boldsymbol{\Pi}_\ell \in \mathbb{R}^{N_\ell \times N_\ell}$, $0 \leq \ell \leq L$, such that $\boldsymbol{\Pi}_0 = \boldsymbol{\Lambda}_0 = \boldsymbol{I}_{N_0}$, $\boldsymbol{\Pi}_L = \boldsymbol{\Lambda}_L = \boldsymbol{I}_{N_L}$, and for every layer $1 \leq \ell \leq L$*

$$\boldsymbol{W}'_\ell = \boldsymbol{\Pi}_\ell \boldsymbol{\Lambda}_\ell \boldsymbol{W}_\ell \boldsymbol{\Lambda}_{\ell-1}^{-1} \boldsymbol{\Pi}_{\ell-1}^{-1} \text{ and } \boldsymbol{b}'_\ell = \boldsymbol{\Pi}_\ell \boldsymbol{\Lambda}_\ell \boldsymbol{b}_\ell. \tag{3}$$

As widely documented [23, 24, 25, 26], PS-equivalent parameters share their realization as proven, e.g., in [9][Lemma 1].

Lemma 1. *For any $\theta, \theta' \in \mathbb{R}^{E \cup \bar{H}}$, if $\theta' \sim_{PS} \theta$ then $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta$.*

A natural question is to determine conditions for the *identifiability* of (the equivalence class up to scaling and permutation of) θ from $\boldsymbol{R}_\theta$. To be more specific, we consider identifiability with respect to a family of parameters Θ, from a set $\mathcal{X}$. A case of particular interest will be when $\mathcal{X}$ is finite, in order to characterize whether θ can be recovered (up to scaling and permutations) from finitely many samples of the network realization $\boldsymbol{R}_\theta$.

Definition 3 (PS-identifiability)**.** *A parameter $\theta \in \Theta \subseteq \mathbb{R}^{E\cup\bar{H}}$ is* PS-identifiable with respect to Θ from $\mathcal{X} \subseteq \mathbb{R}^{N_0}$ *if for every $\theta' \in \Theta$, the equality $\boldsymbol{R}_\theta = \boldsymbol{R}_{\theta'}$ on $\mathcal{X}$ implies $\theta' \sim_{PS} \theta$. When considering $\mathcal{X} = \mathbb{R}^{N_0}$, θ is simply said to be* PS-identifiable *with respect to Θ. When considering $\Theta = \mathbb{R}^{E\cup\bar{H}}$, θ is simply said to be PS-identifiable (from $\mathcal{X}$).*

A trivial observation is that if all outgoing weights of a hidden neuron are zero, then the realization of the network is unchanged under arbitrary modifications of the incoming weights and of the bias of his neuron, hence the corresponding parameter θ cannnot be PS-identifiable with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$. A similar phenomenon occurs if all incoming weights to a hidden neuron are zero. This motivates the definition of admissible parameters and proves Lemma 2 below.

Definition 4. *θ is admissible if for each hidden neuron $\nu \in H$ we have $\boldsymbol{w}_{\bullet\to\nu} \neq 0$ and $\boldsymbol{w}_{\nu\to\bullet} \neq 0$. Equivalently, every hidden neuron belongs to a full path with nonzero weights.*

Lemma 2. *If θ is PS-identifiable with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$ then it is admissible.*

2.4. **An invariant embedding of ReLU networks.** The invariance with respect to (permutations and) scalings (Lemma 1) calls for an invariant representation of equivalence classes of network parameters. A central tool is a representation $\boldsymbol{\Phi}(\theta)$ mapping a network parameter $\theta \in \mathbb{R}^{E\cup\bar{H}}$ to a vector $\boldsymbol{\Phi}(\theta)$ in a space indexed by paths of the network, $\mathbb{R}^{\mathcal{P}}$.

Before going further let us formally introduce paths, as illustrated in Figure 1

Definition 5. *The set $\mathcal{P}_\ell$, $0 \leq \ell \leq L$ (resp. $\mathcal{Q}_\ell$, $1 \leq \ell \leq L-1$) consists of all partial paths from any neuron $\nu_\ell \in N_\ell$ to a neuron of the last (resp.* penultimate*) layer $\nu_L \in N_L$ (resp. $\nu_{L-1} \in N_{L-1}$). Any path $p \in \mathcal{P}_\ell$ is written as a tuple $p = (p_\ell, \dots, p_L)$ where each $p_i \in V$ is a neuron. We say that p is a full path if $\ell = 0$, that is, if p connects the input and the output layers. We may write $p = p_\ell \to p_{\ell+1} \to \cdots \to p_L$, as well as $p = \mu \to q \to \nu$ where $\mu = p_\ell \in N_\ell$, $\nu = p_L \in N_L$ and $q = (p_{\ell+1}, \dots, p_{L-1}) \in \mathcal{Q}_{\ell+1}$.*

We next introduce the representation $\boldsymbol{\Phi}(\cdot)$, which presents some connections with previous work [12, 13, 23] while being more generic as detailed in the introduction.

Definition 6. *Given $\theta \in \mathbb{R}^{E\cup\bar{H}}$, the value of a path is*

$$\Phi_p(\theta) = \Pi_{e\in p}\theta_e, \text{ for each full path } p \in \mathcal{P}_0, \tag{4}$$

$$\Phi_p(\theta) = \theta_{p_\ell}\Pi_{e\in p}\theta_e, \text{for } p = (p_\ell, \dots, p_L) \in \mathcal{P}_\ell, 1 \leq \ell \leq L. \tag{5}$$

For $p \in \mathcal{P}_L$, $p = (\eta)$ with $\eta \in N_L$, $\Phi_p(\theta) = \theta_{p_L} = b_\eta$ is the corresponding output bias. Define $\mathcal{P} := \cup_{\ell=0}^{L}\mathcal{P}_\ell$. For any $\theta \in \mathbb{R}^{E\cup\bar{H}}$ we define

$$\boldsymbol{\Phi}(\theta) := (\Phi_p(\theta))_{p\in\mathcal{P}} \in \mathbb{R}^{\mathcal{P}} \tag{6}$$

Remark 1. *To streamline notations we say that an edge $e = \mu \to \nu \in E$ belongs to p and also write $e \in p$ if there exists $\ell \leq i \leq L-1$ such that $\mu = p_i$ and $\nu = p_{i+1}$. Similarly, we choose to denote $\nu \in p$ if (and only if) the path p* starts *from neuron ν, i.e., when $p = (p_\ell, \dots, p_L) \in \mathcal{P}_\ell$, if $p_\ell = \nu$. With these notations, we can write $\boldsymbol{\Phi}(\theta) = \Pi_{i\in p}\theta_i$.*

This representation characterizes the classes of scaling-equivalent admissible parameters.

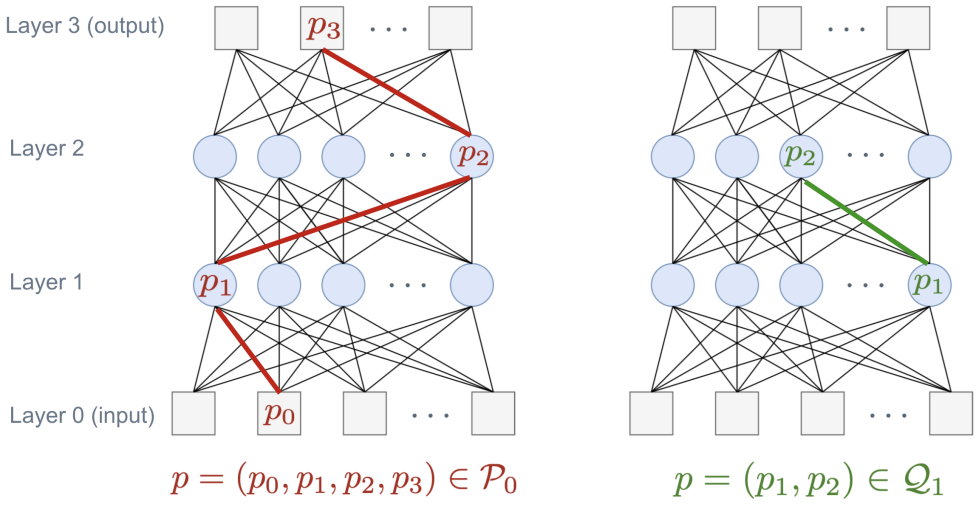

FIGURE 1. We consider a particular network architecture with $L = 4$ layers (equivalently, with two hidden layers). Left: A particular path belonging to $\mathcal{P}_0$. Right: A particular path belonging to $\mathcal{Q}_1$.

Theorem 1. *Consider any* $\theta', \theta \in \mathbb{R}^{E \cup \bar{H}}$.

a) Assume that $\theta \sim_S \theta'$. *Then* $\mathbf{\Phi}(\theta) = \mathbf{\Phi}(\theta')$ *and* $\mathtt{sign}(\theta') = \mathtt{sign}(\theta)$.

b) Assume that θ *is admissible, that* $\mathbf{\Phi}(\theta') = \mathbf{\Phi}(\theta)$, *and that* $\mathtt{sign}(\theta'_E) = \mathtt{sign}(\theta_E)$. *Then* $\theta \sim_S \theta'$ *and* θ' *is also admissible.*

The proof is in Appendix A. A similar result is proven in [27, Theorem 3.3] without considering the biases and by replacing the condition on the signs by a condition on the activation statuses of all partial paths, which depend on the input variable x besides θ.

Remark 2. *The map* $\theta \mapsto \mathbf{\Phi}(\theta)$ *will be referred to as an* embedding *of network parameters. Stricto-sensu, as this map is not an injective function of network parameters, it does not match the definition of an embedding. However, since it characterizes equivalence classes of rescaling-equivalent admissible parameters, it can be used to define without ambiguity an embedding of these equivalence classes in* $\mathbb{R}^{\mathcal{P}}$.

2.5. **Some consequences of PS-identifiability.** Using the embedding $\mathbf{\Phi}(\cdot)$, we show that if θ is PS-identifiable then it is *locally* identifiable *up to scaling only*. Locality is measured in the sense of open balls $B(\boldsymbol{c}, r) = \{\boldsymbol{c}' : \|\boldsymbol{c}' - \boldsymbol{c}\|_\infty < r\}$, where the ambient linear space, equipped with the sup-norm, should always be clear from context.

Definition 7 (local S-identifiability)**.** *Given* $\epsilon > 0$, *a parameter* $\theta \in \Theta \subseteq \mathbb{R}^{E \cup \bar{H}}$ *is* ϵ-locally S-identifiable from $\mathcal{X} \subset \mathbb{R}^{N_0}$ with respect to Θ, *if for every* $\theta' \in \Theta \cap B(\theta, \epsilon)$, *the identity* $\boldsymbol{R}_\theta = \boldsymbol{R}_{\theta'}$ *on* $\mathcal{X}$ *implies* $\theta' \sim_S \theta$. *If there exists* $\epsilon > 0$ *such that* θ *is* ϵ*-locally S-identifiable*

from $\mathcal{X}$ then θ is locally S-identifiable from $\mathcal{X}$. *When $\mathcal{X} = \mathbb{R}^{N_0}$ and/or $\Theta = \mathbb{R}^{E\cup\bar{H}}$ we adopt the same simplified terminology as with the notion of PS-identifiability.*

Remark 3. *If θ is PS-identifiable (resp. locally S-identifiable) from $\mathcal{X} \subseteq \mathbb{R}^{N_0}$ with respect to $\Theta \subseteq \mathbb{R}^{E\cup\bar{H}}$ then the same holds from any $\mathcal{X}' \supseteq \mathcal{X}$ with respect to any $\Theta' \subseteq \Theta$.*

Our first result is the following theorem.

Theorem 2. *Consider $\Theta \subseteq \mathbb{R}^{E\cup\bar{H}}$ and $\mathcal{X} \subset \mathbb{R}^{N_0}$. If $\theta \in \Theta$ is admissible and PS-identifiable from $\mathcal{X}$ with respect to Θ then it is locally S-identifiable from $\mathcal{X}$ with respect to Θ.*

The proof is in Appendix B and uses the embedding $\mathbf{\Phi}(\cdot)$. By Lemma 2, PS-identifiability with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$ implies admissibility. Considering any Θ with a similar property, a direct corollary of Theorem 2 is that PS-identifiability with respect to Θ implies local S-identifiability with respect to Θ.

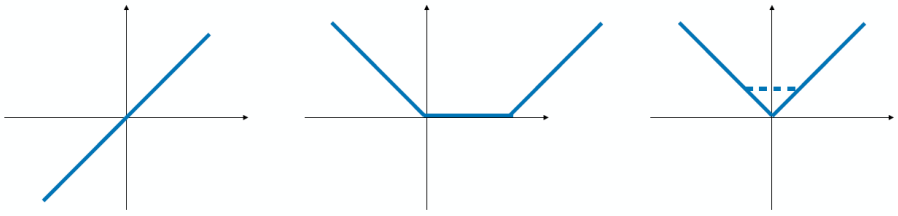

FIGURE 2. Realizations of networks from (a) Example 1; (b) Example 2 ; (c) Example 4

An example shows that indeed, local S-identifiability depends on the constraint set Θ.

Example 1 (see Figure 2-(a))**.** *On a shallow network architecture with two hidden neurons ν_1, ν_2, the identity $\mathtt{id} : \mathbb{R} \to \mathbb{R}, x \mapsto x$ can be written as $x = \mathtt{ReLU}(x-t) - \mathtt{ReLU}(-(x-t)) + t = \boldsymbol{R}_{\theta_t}$ with $\theta_t = (w_{\mu\to\nu_1} = 1, w_{\mu\to\nu_2} = -1, b_{\nu_1} = -t, b_{\nu_2} = t, w_{\nu_1\to\eta} = 1, w_{\nu_2\to\eta} = -1, b_\eta = t)$ for every $t \in \mathbb{R}$ (μ is the input neuron, η the output neuron). Since θ_t and $\theta_{t'}$, $t \neq t'$ have different output bias, they are not PS-equivalent. This shows that, e.g., θ_0 is not locally S-identifiable with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$. With respect to the set Θ of networks without output bias ($b_\eta = 0$), as detailed in Example 5, θ_0 becomes PS-identifiable from $\mathcal{X} = \mathbb{R}$.*

The above example includes two neurons which are *twins* in the following sense.

Definition 8 (Twin neurons)**.** *Consider a parameter θ on a network architecture of any depth. Two hidden neurons $\nu \neq \nu'$ from the same layer are said to be twins if there exists $\lambda \in \mathbb{R}$ such that $(\boldsymbol{w}_{\bullet\to\nu}, b_\nu) = \lambda(\boldsymbol{w}_{\bullet\to\nu'}, b_{\nu'})$. If θ is admissible then necessarily $\lambda \neq 0$, and ν, ν' are said to be positive twins if $\lambda > 0$, negative twins otherwise.*
NB: Even though each hidden neuron $\nu \in H$ is (positive) twin to itself, such a neuron is abusively said to have "no twin" if it is not twin with any $\nu' \neq \nu$ from the same layer. We also say that θ has no twins if none of its neurons have any twin.

Intuitively, if ν, ν' are twins then the corresponding pre-activation functions $z_\nu(\theta, \cdot)$ $z_{\nu'}(\theta, \cdot)$ are collinear, and the resulting post-activation functions, $y_\nu(\theta, \cdot)$ $y_{\nu'}(\theta, \cdot)$ are also collinear for positive twins. For negative twins, there exists linear combinations of the post-activations that are simply proportional to the pre-activations, somehow bypassing the effect of the ReLU nonlinearity. As proved in Appendix C, twins *always* prevent identifiability with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$.

Lemma 3. *Consider $\theta \in \Theta = \mathbb{R}^{E \cup \bar{H}}$.*

a) Assume that θ is locally S-identifiable with respect to Θ. Then θ has no positive twins.

b) Assume that θ is PS-identifiable from some bounded set $\mathcal{X} \subseteq \mathbb{R}^{N_0}$ with respect to Θ. Then θ has no twins.

We will see in Example 4 (in section 5) that the absolute value function (see Figure 2-(c)) is the realization of a shallow network with two hidden neurons that are negative twins, yet it is PS-identifiable (hence locally S-identifiable) with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$. It is even locally S-identifiable from some finite set $F \subseteq \mathbb{R}$. Of course, by Lemma 3 such a network cannot be PS-identifiable from any bounded set with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$.

Twins are a form of *local* degeneracy. For shallow networks, we will show that this is the only form of local degeneracy (see the upcoming Lemma 5 and Theorem 3), but we will see other forms for deeper networks (see Example 3). As illustrated next, there are also *non-local* degeneracies that can prevent identifiability.

Example 2 (see Figure 2-(b))**.** *The function*

$$f(x) = \begin{cases} -x, & \text{if } x \leq 0 \\ 0, & \text{if } 0 \leq x \leq 1 \\ x - 1, & \text{if } x \geq 1 \end{cases}$$

satisfies $f(x) = \texttt{ReLU}(-x) + \texttt{ReLU}(x-1) = \texttt{ReLU}(x) + \texttt{ReLU}(-(x-1)) - 1$. It is thus the realization of $\theta = (w_{\mu\to\nu_1} = -1, w_{\mu\to\nu_2} = 1, b_{\nu_1} = 0, b_{\nu_2} = -1, w_{\nu_1\to\eta} = w_{\nu_2\to\eta} = 1, b_\eta = 0$, but also of $\theta' = (w'_{\mu\to\nu_1} = 1, w'_{\mu\to\nu_2} = -1, b'_{\nu_1} = 0, b'_{\nu_2} = 1, w'_{\nu_1\to\eta} = w'_{\nu_2\to\eta} = 1, b'_\eta = -1$, which are not PS-equivalent since $b_\eta \neq b'_\eta$. Yet the theory we establish (see Lemma 5) shows that θ and θ' are both locally S-identifiable from some finite set $F \subset \mathbb{R}$.

It turns out that the above example fails to be *irreducible* as we formalize next.

Definition 9 (Irreducibility)**.** *A parameter θ is irreducible if for each hidden layer $1 \leq \ell \leq L-1$ and non-empty subset $T \subset N_\ell$ we have*

$$\boldsymbol{W}_{\ell+1} \boldsymbol{I}_T \boldsymbol{W}_\ell \neq 0, \quad \text{with } \boldsymbol{I}_T = \texttt{diag}(\boldsymbol{\chi}_T), \tag{7}$$

with $\boldsymbol{\chi}_T \in \{0,1\}^{N_\ell}$ the indicator function of T: $(\boldsymbol{\chi}_T)_\nu = 1$ if, and only if, $\nu \in T$. We denote $\Theta_{\texttt{irr}} \subset \mathbb{R}^{E \cup \bar{H}}$ the set of all irreducible parameters.

Fact 3. *Each irreducible parameter is also admissible.*

In fact, as established in Appendix D, any PS-identifiable parameter with no twins must be irreducible.

Lemma 4. *If θ is PS-identifiable from $\mathcal{X} \subseteq \mathbb{R}^{N_0}$ with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$ and has no twin, then it is irreducible.*

In particular, in light of Lemma 3, every PS-identifiable parameter is irreducible. In the shallow case, a direct consequence of irreducibility can be obtained using an "algebraic" expression of the realization $\boldsymbol{R}_\theta$ (Lemma 9 in section 4): for every input vector x where $\boldsymbol{R}_\theta$ is differentiable, the Jacobian of $\boldsymbol{R}_\theta$ is given by $\boldsymbol{W}_2 \boldsymbol{I}_1 \boldsymbol{W}_1$ with $\boldsymbol{I}_1 = \mathtt{diag}(\boldsymbol{a}_1(\theta, x))$. Irreducibility thus implies that this Jacobian can only vanish if $\boldsymbol{a}_1(\theta, x) = \boldsymbol{0}$, i.e., if all neurons are inactive. As illustrated on Example 2 (see Figure 2-(b)) this however does not *characterize* irreducibility, and an intuitive characterization of irreducibility in terms of simple properties of $\boldsymbol{R}_\theta$ is left to future work.

2.6. **Identifiability conditions in the shallow case.** For shallow neural networks, we prove that admissible parameters with no twins are locally S-identifiable from a *finite* set. Such results resonate with previous work on the identifiability of shallow networks equipped with various activation functions other than the ReLU [4, 5, 6, 7].

Lemma 5. *Consider a shallow architecture. If θ is admissible with no twins, then there is a finite $\mathcal{X} \subseteq \mathbb{R}^{N_0}$ with $\mathtt{card}(\mathcal{X}) \leq (|N_0|+1)(|N_1|+1)$ from which θ is locally S-identifiable with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$.*

The proof is in Section 5.2. Combined with irreducibility, the absence of twins is further shown to be equivalent to PS-identifiability from a *bounded* set. Whether this is also equivalent to PS-identifiability from a *finite* set is left to future work, as well as a possible explicit control of the cardinality of such a finite set.

Theorem 3. *Consider a shallow network architecture. The following are equivalent*
a) there is a bounded $\mathcal{X} \subseteq \mathbb{R}^{N_0}$ from which θ is PS-identifiable with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$;
b) θ has no twins and is irreducible.

Proof. The implication a $\Rightarrow$ b is a consequence of Lemma 3 and Lemma 4. The converse b $\Rightarrow$ a follows by Theorem 6-b in Section 5.3. □

As established with Theorem 6-a in Section 5.3, the architecture itself is identifiable in the following sense for irreducible parameters with no twins.

Theorem 4. *Consider two shallow network architectures with the same input and output layers, N_0 and N_2, and potentially distinct hidden layer $H = N_1$, $H' = N_1'$. Let θ, θ' be parameters on each architecture. Assume that θ is irreducible with no twins, and that θ' is admissible with no twins. If $\boldsymbol{R}_\theta = \boldsymbol{R}_{\theta'}$ on $\mathbb{R}^{N_0}$ then $\mathtt{card}(N_1) = \mathtt{card}(N_1')$ and $\theta' \sim_{PS} \theta$.*

As illustrated by Example 4 in section 5, there are also shallow networks that are PS-identifiable from $\mathcal{X} = \mathbb{R}^{N_0}$ but not from any bounded set. They are of course irreducible by Lemma 4, and have no positive twin by Lemma 3, but they have one or more pair of negative twins.

2.7. **A glimpse at the analysis of local identifiability.** Much of the local identifiability analysis, which is conducted in details in section 4, relies on an important property of the embedding $\boldsymbol{\Phi}$ (besides its ability to characterize scaling equivalence, see Theorem 1): it provides a *locally linear parameterization* of the realization of the network, in the sense that given θ and for "most" $x \in \mathbb{R}^{N_0}$ we have, for every θ' in a (small enough) neighborhood θ

$$\boldsymbol{R}_{\theta'}(x) - \boldsymbol{R}_{\theta}(x) = \boldsymbol{C}_{\theta,x} \cdot \big(\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta)\big) \tag{8}$$

with $\boldsymbol{C}_{\theta,x} \in \mathbb{R}^{N_L \times \mathcal{P}}$ some linear operator that is *independent of* θ'. This property holds provided x is a point where the gradient of $\boldsymbol{R}_\theta$ (and of all pre-activations at intermediate hidden layers) is well-defined and continuous, which motivates the following definition.

Definition 10. *Consider any network architecture. Given a parameter θ we define for each hidden neuron $\nu \in H$ the set $\Gamma_\nu(\theta)$ of input vectors where $z_\nu(\theta, x) = 0$ and the gradient $\nabla z_\nu(\theta, x)$ is* well-defined *and nonzero,*

$$\Gamma_\nu(\theta) := \{x \in \mathbb{R}^{N_0} : z_\nu(\theta, x) = 0 \text{ and } \nabla z_\nu(\theta, x) \neq 0\}. \tag{9}$$

We define $\mathcal{X}_\theta \subseteq \mathbb{R}^{N_0}$ as the complement to $\cup_{\nu \in H} \Gamma_\nu(\theta)$.

Definition 10 is extremely close to the definition of *Bent Hyperplanes* [28] (except that we add the non-nullity condition on the gradient). Informally, and as previously stated [29, 30, 31], bent hyperplanes separate the input space into *linear regions* where the realization of the network $x \mapsto \boldsymbol{R}_\theta(x)$ is affine, see Figure 3 for an illustration.

For our needs, we will provide in Lemma 11 an alternate characterization of $\mathcal{X}_\theta$ which we have not found elsewhere in the literature. It will be used in Corollary 3 to formalize Property (8) for $x \in \mathcal{X}_\theta$, which motivates the definition of non-degenerate parameters.

Definition 11 (Non-degeneracy)**.** *Consider the finite dimensional linear space*

$$\mathtt{V}(\theta) := \cap_{x \in \mathcal{X}_\theta} \mathtt{ker}(\boldsymbol{C}_{\theta,x}) \subseteq \mathbb{R}^{\mathcal{P}}. \tag{10}$$

A parameter $\theta \in \Theta \subseteq \mathbb{R}^{E \cup \bar{H}}$ is ϵ-non-degenerate *with respect to Θ, where $\epsilon > 0$, if it is admissible and for every $\theta' \in \Theta \cap B(\theta, \epsilon)$ we have*

$$\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta) \Rightarrow \boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta). \tag{11}$$

It is non-degenerate *with respect to Θ if there exists $\epsilon > 0$ such that it is ϵ-non-degenerate with respect to Θ.*

Exploiting the fact that all considered spaces are finite dimensional, we characterize the space $\mathtt{V}(\theta)$ in terms of certain *activation spaces* (Definition 14) and prove that non-degeneracy is equivalent (see Theorem 5, the main result of section 4) to the existence of some *finite set $F \subset \mathcal{X}_\theta$* such that θ is locally S-identifiable from F (hence also locally S-identifiable from $\mathcal{X} = \mathbb{R}^{N_0}$). The cardinality of F is bounded from above using the dimension of activation spaces.

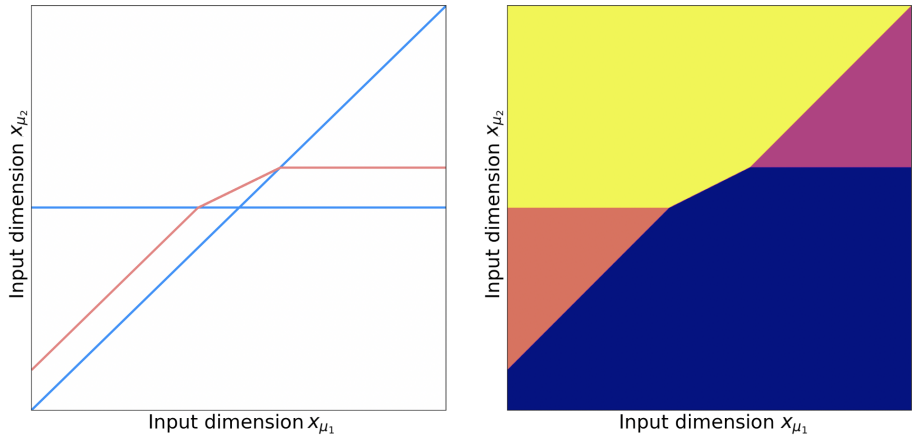

FIGURE 3. We consider a network architecture with $|N_1| = 2$ neurons on the first hidden layer and $|N_2| = 1$ neuron on the second hidden layer. The input x is two-dimensional: $|N_0| = 2$ and the output is scalar: $|N_4| = 1$. Left: bent hyperplanes for the first hidden layer, $\Gamma_\nu(\theta)$, $\nu \in N_1$ (blue) and second hidden layer $\Gamma_\nu(\theta)$, $\nu \in N_2$ (red). Right: linear regions. All the weights and biases were initialized randomly. The figures are generated with a PyTorch script available at `https://github.com/pierrestock/linear-regions/blob/main/partition.ipynb`.

2.8. **Non-degeneracy and irreducibility in shallow *vs* deeper architectures.** An easy sufficient condition for non-degeneracy is to have a trivial space $\mathtt{V}(\theta) = \{0\}$. For *scalar-valued* shallow networks ($L = 2, |N_L| = 1$), we prove (cf Lemma 17 and Corollary 2 that non-degeneracy with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$ is in fact *equivalent* to $\mathtt{V}(\theta) = \{0\}$, and for shallow (possibly vector-valued) networks, the latter is proved to hold if, and only if, there are no twins (by Corollary 2 and Lemma 15). In light of Theorem 3, when combined with irreducibility, the fact that $\mathtt{V}(\theta) = \{0\}$ thus becomes equivalent (for shallow networks) to the PS-identifiability of θ from some bounded set.

For networks of depth $L \geq 3$, any parameter such that $\mathtt{V}(\theta) = \{0\}$ is of course still non-degenerate (hence locally S-identifiable from a finite set, by Theorem 5), but this property is no longer equivalent to the absence of twins: further conditions between layers are required, as illustrated by the following example.

Example 3. *In Figure 4, we exhibit a two-hidden-layer architecture valued with a parameter θ that presents no twin neurons (see Definition 8) but such that θ is not locally S-identifiable (see Definition 7).*

Characterizing concrete conditions ensuring $\mathtt{V}(\theta) = \{0\}$ is left to future work. A particular challenge is to understand whether the condition $\mathtt{V}(\theta) = \{0\}$, combined with (a possibly strengthened version of) irreducibility remains equivalent to PS-identifiability from a bounded set. We note that irreducibility in the shallow case is reminiscent of [32, Equation

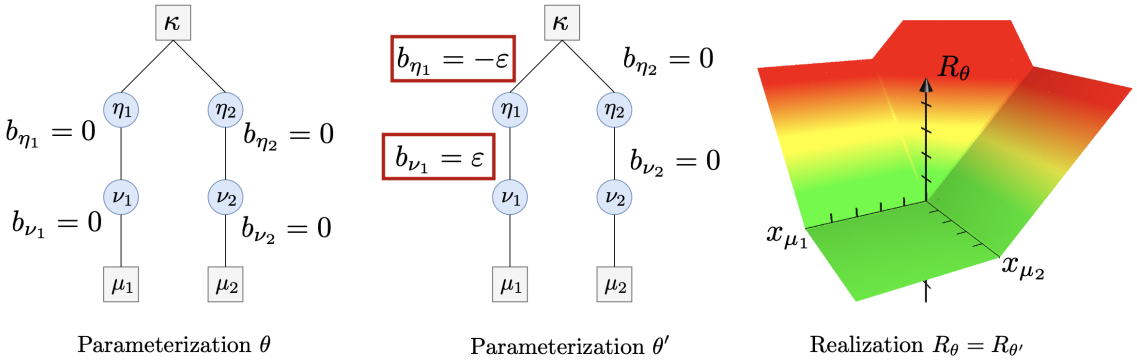

FIGURE 4. A network with two hidden layers that is not locally S-identifiable, while having no twin hidden neurons. Weights w_e and w'_e, $e \in E$ are set to one on the displayed edges and to zero on other edges, and are not depicted here for readability. Left: parameterization θ valuing the architecture. Center: alternative parameterization θ' such that θ and θ' are *not* rescaling equivalent. Right: the two realizations R_θ and $R_{\theta'}$ coincide: for every input point $x = (x_{\mu_1}, x_{\mu_2}) \in \mathbb{R}^2, R_\theta(x) = R_{\theta'}(x)$. The construction is valid for arbitrary $\varepsilon > 0$.

(8)], a condition used to define so-called "general ReLU networks" to provide sufficient identifiability conditions in deeper settings. This may serve as a guide to identify stronger notions of irreducibility for deep networks. Preliminary investigations suggest that certain tensor products of activation vectors play a role when analyzing non-degeneracy. This is reminiscent of the tools studied by Fornasier *et al.* [10] with two hidden layers $L = 3$ in a smooth context that cannot cover ReLU networks.

2.9. **Discussion.** Before diving into the technical contributions in the next Sections, we discuss some topics of interest for the reader that are mostly out of the scope of this work. We refer the reader to Figure 5 for a brief summary of the results proven in this paper.

Local identifiability and optimization. First, we argue that studying *local* (instead of global) S-identifiability is of practical interest, as discussed e.g. in [12, 13]. Indeed, neural networks are traditionally optimized with a variant of stochastic gradient descent, or SGD [33]. Hence, (1) during training, the optimization yields parameters that are close to the previous ones and (2) the parameters obtained after convergence can be expected to be locally optimal up to natural permutation and rescaling equivalences.

Identifiability from a finite *set.* Since we are mainly interested in the problem of recovering (the equivalence class of) θ from the knowledge of its realization $\boldsymbol{R}_\theta$, we list below some questions calling for extensions of Theorem 3. Indeed, it is not always possible to recover θ from its realization. Even when such a recovery is theoretically possible, it may

involve having full access to the function $\boldsymbol{R}_\theta$, which is not a concrete input to provide to any reconstruction algorithm. A more practical question is: when can we recover (the equivalence class of) θ from the knowledge of *finitely many samples* $\boldsymbol{R}_\theta(x_i), 1 \leq i \leq n$? When there exists a choice (that may depend on θ) of n and x_i, $1 \leq i \leq n$ such that this is feasible, we also get as a byproduct a reconstruction of $\boldsymbol{R}_\theta$ from the sole knowledge of its samples at these points. Hence, another question of interest is: when can the function $\boldsymbol{R}_\theta$ be identified from the knowledge of finitely many of its samples ? This is possibly less demanding, as here it is not required to be able to reconstruct (the equivalence class of) θ from its realization. In both cases, since θ is not known beforehand, it is important to ensure that the choice of the sampling set is algorithmically feasible, for example if it is done iteratively at least the first sample must be chosen without any knowledge on θ or $\boldsymbol{R}_\theta$. Of course, answers to these questions lead to further ones, that we do not touch upon: if θ can be identified from finitely many samples, how many samples are sufficient[4] (resp. necessary) ? Can we explicit an scheme (possibly randomized) to choose these samples ? Can we explicit an algorithm to perform reconstruction ? How stable is it to inaccuracies in the evaluation of $\boldsymbol{R}_\theta(x_i)$ or to the knowledge of x_i?

Reverse-engineering ReLU networks. Here, we dicuss the work of Rolnick and Kording [9] more extensively than what was done in the Introduction. The goal is to position our work with respect to this interesting work. The authors present a sampling algorithm to recover a ReLU network's architecture and parameters, up to permutations and rescalings. The authors prove that their algorithm terminates except for a measure-zero set of networks and do not provide the complexity of their method in terms of number of the samples needed to recover $\boldsymbol{R}_\theta$, except for recovering the first layer's parameters. They reason in terms of so-called *activation* and *linear* regions [28] and make the following assumptions. Recall that the sets $\Gamma_\nu(\theta)$ are introduced in Definition 9 for every hidden neuron ν. $\Gamma_\nu(\theta)$ is often called the *separating* or *bent* hyperplane for neuron ν.

(1) Linear Regions assumption as stated by the authors: "Each [activation][5]region represents a maximal connected component of input space on which the [realization $\boldsymbol{R}_\theta$] is given by a single linear function". In other words, the authors assume that activation regions and linear regions coincide (Section 3.2 in the original paper).
(2) All the sets $\Gamma_\nu(\theta)$ for $\nu \in H$ have codimension 1[6] hence the name *separating hyperplane* (implicitly assumed, see in particular the first paragraph of Section 3.3).
(3) For every hidden neuron ν in layer $1 \leq \ell \leq L-1$, $\Gamma_\nu(\theta)$ intersects *all* the sets $\Gamma_{\nu'}(\theta)$ for all neurons ν' in a previous layer $1 \leq \ell' < \ell \leq L-1$ (Section 5.2 in the original paper).
(4) For $\nu \neq \nu'$ such that ν belong to layer ℓ and ν' belongs to layer $\ell' < \ell$, "$\Gamma_\nu(\theta)$ bends on $\Gamma_{\nu'}(\theta)$, but $\Gamma_\nu(\theta)$ and $\Gamma_{\nu'}(\theta)$ cannot both bend at their intersection" (implicitly assumed, see in particular the first paragraph of Section 3.3).

[4]Lemma 5 partly answers this question regarding local S-identifiability for shallow networks.
[5]What the authors denote as linear regions are in fact known as activation regions, see [28].
[6]This prevents the case where $\nabla R_\theta(x) = 0$ for $x \in B(x_0, r)$

(5) For every hidden neuron $\nu \in H$, $\Gamma_\nu(\theta)$ is *not bounded* and *not disconnected* (Section 5.2 in the original paper).

According to the authors, parameters θ that do not satisfy at least one of these assumptions constitute a measure-zero set of networks, hence the authors discard these cases from their analysis. In the remainder of this paper, we aim at more precisely characterizing this measure-null zero set. This is fully done in the shallow case, and the developed tools should be instrumental when pursuing this mathematical study in deeper settings.

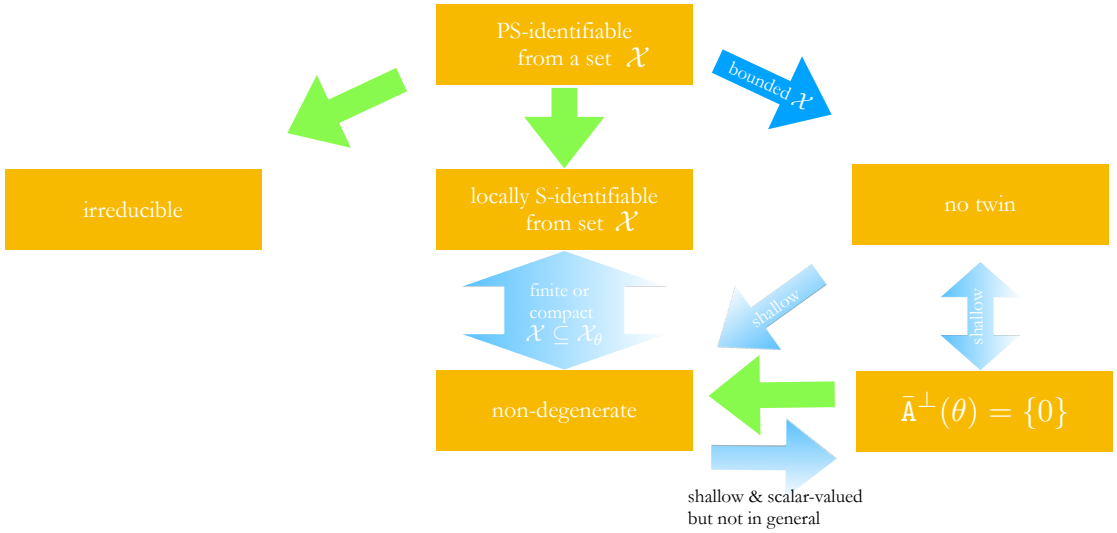

FIGURE 5. Summary of the various results proven in the paper. The space $\bar{\mathtt{A}}^\perp(\theta)$ is defined in Definition 14. Theorem 3 further establishes that irreducibility and the absence of twins imply PS-identifiability from a bounded set in the shallow case.

3. Rescaling invariance of the embedding

The proof of the main property of the embedding $\mathbf{\Phi}(\cdot)$, Theorem 1, exploits linear operators related to $\mathbf{\Phi}(\cdot)$. The following definition is motivated by the obvious observation that, if θ has positive entries $\theta_i = e^{\alpha_i}$, $i \in E \cup \bar{H}$, then $\mathbf{\Phi}(\theta) = e^{\boldsymbol{P}\alpha}$ where the exponential is taken componentwise. This is related to the idea of updating weights *multiplicatively*, which is exploited in particular by Bernstein [34] to investigate learning stability.

Definition 12. *Consider* $\boldsymbol{P} : \mathbb{R}^{E\cup\bar{H}} \to \mathbb{R}^{\mathcal{P}}$ *the linear operator defined for* $\boldsymbol{u} \in \mathbb{R}^{E\cup\bar{H}}$ *as*

$$(\boldsymbol{P}\boldsymbol{u})_p := \begin{cases} \sum_{e\in p} u_e, & \text{for each full path } p \in \mathcal{P}_0; \\ u_{p_\ell} + \sum_{e\in p} u_e, & \text{for each partial path } p = (p_\ell, \dots, p_L) \in \mathcal{P}_\ell,\ 1 \le \ell \le L. \end{cases} \tag{12}$$

With the notations from Remark 1 we can also write $(\boldsymbol{P}\boldsymbol{u})_p = \sum_{i\in p} u_i$.

Before proving Theorem 1 we express a few technical lemmas.

Lemma 6. *For every $\theta \in \mathbb{R}^{E\cup\bar{H}}$, with $\mathtt{supp}(\mathbf{\Phi}(\theta)) = \{p \in \mathcal{P} : \Phi_p(\theta) \neq 0\}$ we have*

$$\{i \in E \cup \bar{H} : \exists p \in \mathtt{supp}(\mathbf{\Phi}(\theta)), p \ni i\} \subseteq \{i \in E \cup \bar{H} : \theta_i \neq 0\} = \mathtt{supp}(\theta). \tag{13}$$

If $\theta \in \mathbb{R}^{E\cup\bar{H}}$ is admissible then we further have

$$\mathtt{supp}(\theta) = \{i \in E \cup \bar{H} : \exists p \in \mathtt{supp}(\mathbf{\Phi}(\theta)), p \ni i\}. \tag{14}$$

Proof. For each path $p \in \mathcal{P}$ denote $I_p = \{i \in E \cup \bar{H} : i \in p\}$ and observe first that the left hand side in (13) is $\cup_{p\in\mathtt{supp}(\mathbf{\Phi}(\theta))} I_p$. Consider $p \in \mathtt{supp}(\mathbf{\Phi}(\theta))$. Since $\Phi_p(\theta) = \Pi_{i\in p}\theta_i$, we have $\theta_i \neq 0$ for each $i \in I_p$, i.e., $I_p \subseteq \mathtt{supp}(\theta)$. As this holds for every $p \in \mathtt{supp}(\mathbf{\Phi}(\theta))$ we obtain $\cup_{p\in\mathtt{supp}(\mathbf{\Phi}(\theta))} I_p \subseteq \mathtt{supp}(\theta)$. This establishes (13).

Assuming now that θ is admissible, consider $i \in \mathtt{supp}(\theta)$ and distinguish three cases. If $i = \eta \in N_L$ is an output neuron, then $p = (\eta) \ni i$ yields $\Phi_p(\theta) = \theta_\eta \neq 0$. If $i = \nu \in H$ is a hidden neuron, then since θ is admissible there is a path $p \ni i$ with nonzero weights connecting ν to an output neuron. This path satisfies $\Phi_p(\theta) \neq 0$. Finally, if $i = \nu \to \nu'$ is an edge, then since θ is admissible there is a path connecting the input layer to ν and a path connecting ν' to the output layer, both with nonzero weights. Concatenating them yields a path $p \ni i$ such that $\Phi_p(\theta) \neq 0$. In all cases, we obtain the existence of a path $p \in \mathtt{supp}(\mathbf{\Phi}(\theta))$ such that $p \ni i$. This establishes (14). □

Corollary 1. *Consider $\theta, \theta' \in \mathbb{R}^{E\cup\bar{H}}$ such that $\mathbf{\Phi}(\theta') = \mathbf{\Phi}(\theta)$. If θ is admissible then $\mathtt{supp}(\theta') = \mathtt{supp}(\theta)$ and θ' is also admissible.*

Proof. By Lemma 6 and the equality $\mathbf{\Phi}(\theta') = \mathbf{\Phi}(\theta)$ we have

$$\begin{aligned}\mathtt{supp}(\theta) &= \{i \in E \cup \bar{H} : \exists p \in \mathcal{P}, \Phi_p(\theta) \neq 0, i \in p\} \\ &= \{i \in E \cup \bar{H} : \exists p \in \mathcal{P}, \Phi_p(\theta') \neq 0, i \in p\} \subseteq \mathtt{supp}(\theta').\end{aligned}$$

The fact that θ is admissible is a property of its support, and the inclusion $\mathtt{supp}(\theta) \subseteq \mathtt{supp}(\theta')$ implies that θ' is also admissible. It follows using Lemma 6 again that the rightmost inclusion above is an equality. □

Lemma 7. *Given $\theta \in \mathbb{R}^{E\cup\bar{H}}$ an admissible parameter, consider the spaces*

$$W_\theta := \{\alpha \in \mathbb{R}^{E\cup\bar{H}}, [\mathbf{\Phi}(\theta) \odot \boldsymbol{P}\alpha]_{\mathcal{P}_0} = 0\} \tag{15}$$

$$V_\theta := \{\alpha \in W_\theta, \alpha_{\bar{H}} = 0, \mathtt{supp}(\alpha) \subseteq \mathtt{supp}(\theta)\}. \tag{16}$$

Given $\alpha \in W_\theta$, define for each hidden neuron $\nu \in H$

$$(\boldsymbol{S}_\theta\alpha)_\nu \triangleq -\sum_{e\in p} \alpha_e \tag{17}$$

with p any path with edges $e \in E \cap \mathtt{supp}(\theta)$ joining ν to an output neuron η.

a) The linear map $\boldsymbol{S}_\theta \colon W_\theta \to \mathbb{R}^H$ is well-defined and independent of the choice of p and η;

b) Its restriction $\boldsymbol{S}_\theta : V_\theta \to \mathbb{R}^H$ *is an isomorphism. Its inverse* $\boldsymbol{S}_\theta^{-1} : \mathbb{R}^H \to V_\theta$ *is such that for any* $\beta \in \mathbb{R}^H$, $\boldsymbol{S}_\theta^{-1}\beta = \alpha$ *where* $\alpha_{\bar{H}} = 0$ *and for each edge* $e = \mu \to \nu \in E \cap \mathtt{supp}(\theta)$,

$$\alpha_e \triangleq \begin{cases} -\beta_\mu & \text{if } \mu \in N_{L-1} \text{ (and } \nu \in N_L) \\ \beta_\nu - \beta_\mu & \text{if } \mu \in N_\ell, 1 \leq \ell \leq L-2 \\ \beta_\nu & \text{if } \mu \in N_0. \end{cases} \tag{18}$$

while $\alpha_e = 0$ *for each* $e \in E \backslash \mathtt{supp}(\theta)$.

The proof is postponed to Appendix A to keep the reading flow.

Proof of Theorem 1. By Definition 1, $\theta \sim_S \theta'$ if, and only if, there are $\{\lambda_\nu\}_{\nu \in H \cup N_0 \cup N_L}$ such that

$$\lambda_\nu > 0, \quad \forall \nu \in H, \quad \text{and } \lambda_\nu = 1, \quad \forall \nu \in N_0 \cup N_L \tag{19}$$

$$\theta'_e = \lambda_\mu^{-1} \theta_e \lambda_\nu, \quad \forall e = \mu \to \nu \in E \quad \text{and } \theta'_\nu = \theta_\nu \lambda_\nu, \quad \forall \nu \in \bar{H}. \tag{20}$$

Thus, if $\theta' \sim_S \theta$ then $\mathtt{sign}(\theta') = \mathtt{sign}(\theta)$, and for every path $p = (p_0, \ldots, p_L) \in \mathcal{P}_0$ we get

$$\Phi_p(\theta') = \Pi_{k=0}^{L-1} \theta'_{p_k \to p_{k+1}} = \Pi_{k=0}^{L-1} (\lambda_{p_k}^{-1} \theta_{p_k \to p_{k+1}} \lambda_{p_{k+1}}) = \Pi_{k=0}^{L-1} \theta_{p_k \to p_{k+1}} = \Phi_p(\theta),$$

while for $p = (p_\ell, \ldots, p_L) \in \mathcal{P}_\ell$, $1 \leq \ell \leq L$

$$\Phi_p(\theta') = \theta'_{p_\ell} \Pi_{k=\ell}^{L-1} \theta'_{p_k \to p_{k+1}} = \theta_{p_\ell} \lambda_{p_\ell} \Pi_{k=\ell}^{L-1} (\lambda_{p_k}^{-1} \theta_{p_k \to p_{k+1}} \lambda_{p_{k+1}}) = \theta_{p_\ell} \Pi_{k=0}^{L-1} \theta_{p_k \to p_{k+1}} = \Phi_p(\theta).$$

This shows $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$.

Conversely, assume that θ is admissible and that $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$ and $\mathtt{sign}(\theta'_E) = \mathtt{sign}(\theta_E)$. By Corollary 1, since θ is admissible and $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$, we have $\mathtt{supp}(\theta') = \mathtt{supp}(\theta)$ hence there are $\gamma_i \neq 0, i \in \mathtt{supp}(\theta)$ such that $\theta'_i = \gamma_i \theta_i$ for each $i \in \mathtt{supp}(\theta)$. Since $\mathtt{sign}(\theta'_E) = \mathtt{sign}(\theta_E)$, we have $\gamma_e > 0$ for every $e \in E \cap \mathtt{supp}(\theta)$. Consider $\alpha \in \mathbb{R}^{E \cup \bar{H}}$ such that $\alpha_{\bar{H}} = 0$, $e^{\alpha_e} = \gamma_e$ for $e \in E \cap \mathtt{supp}(\theta)$, and $\alpha_e = 0$ for $e \in E \backslash \mathtt{supp}(\theta)$. For each $p \in \mathcal{P}_0$

$$\Phi_p(\theta') = \Pi_{e \in p} \theta'_e = \Pi_{e \in p} (\theta_e e^{\alpha_e}) = \Phi_p(\theta) e^{\sum_{e \in p} \alpha_e} = \Phi_p(\theta) \odot e^{(\boldsymbol{P}\alpha)_p}.$$

Since $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$, it follows that for each $p \in \mathcal{P}_0$ such that $\Phi_p(\theta) \neq 0$ we have $e^{(\boldsymbol{P}\gamma)_p} = 1$, i.e., $(\boldsymbol{P}\alpha)_p = 0$. Thus, $\Phi_p(\theta)(\boldsymbol{P}\alpha)_p = 0$ for all $p \in \mathcal{P}_0$, i.e., $[\boldsymbol{\Phi}(\theta) \odot \boldsymbol{P}\alpha]_{\mathcal{P}_0} = 0$. Since $\alpha_{\bar{H}} = 0$, we get that α belongs to the space V_θ defined in (16) in Lemma 7. Since θ is admissible, the linear operator $\boldsymbol{S}_\theta$ defined in Lemma 7 is a well-defined bijection from V_θ to $\mathbb{R}^H$, hence α is related to $\beta := \boldsymbol{S}_\theta \alpha \in \mathbb{R}^H$ through the relation (18). Considering $\delta \in \mathbb{R}^{N_0 \cup H \cup N_L}$ with $\delta_\nu := \beta_\nu$ for $\nu \in H$, $\delta_\nu = 0$ for $\nu \in N_0 \cup N_L$, relation (18) implies

$$\alpha_e = \delta_\nu - \delta_\mu, \quad \forall e = \mu \to \nu \in E \cap \mathtt{supp}(\theta).$$

Setting $\lambda_\nu := e^{\delta_\nu}$ for each $\nu \in N_0 \cup H \cup N_L$, it follows that for each $e = \mu \to \nu \in E$ we have $\theta'_e = \lambda_\mu^{-1} \theta_e \lambda_\nu$. Since $\mathtt{supp}(\theta) = \mathtt{supp}(\theta)$, this also trivially holds for $e \in E \backslash \mathtt{supp}(\theta)$.

To conclude, we show that $\theta'_\nu = \theta_\nu \lambda_\nu$ for each $\nu \in \bar{H}$. As this holds trivially for $\nu \in \bar{H} \cap \mathtt{supp}(\theta)$, we focus on $\nu \in \bar{H} \cap \mathtt{supp}(\theta)$. First, we treat the case of $\eta \in N_L \cap \mathtt{supp}(\theta)$ by observing that, with $p = (\eta) \in \mathcal{P}_L$ we have $\theta'_\eta = \Phi_p(\theta') = \Phi_p(\theta) = \theta_\eta = \theta_\eta \lambda_\eta$ since $\lambda_\eta = e^{\delta_\eta} = 1$ by definition of $\delta_\eta := 0$. Now consider $\nu \in H \cap \mathtt{supp}(\theta)$. Since θ is admissible,

there is a partial path p connecting ν to some output neuron η with edges in $\mathtt{supp}(\theta)$. Since $-\sum_{e\in p}\alpha_e = (\boldsymbol{S}_\theta\alpha)_\nu = \beta_\nu = \delta_\nu$ we have

$$\begin{aligned}\theta'_\nu\Pi_{e\in p}\theta'_e = \Phi_p(\theta') = \Phi_p(\theta) = \theta_\nu\Pi_{e\in p}\theta_e = \theta_\nu\Pi_{e\in p}\theta'_e e^{-\alpha_e} &= \theta_\nu(\Pi_{e\in p}\theta'_e)e^{-\sum_{e\in p}\alpha_e}\\ &= \theta_\nu(\Pi_{e\in p}\theta'_e)e^{\delta_\nu} = \theta_\nu(\Pi_{e\in p}\theta'_e)\lambda_\nu.\end{aligned}$$

We conclude using that $\Pi_{e\in p}\theta'_e \neq 0$ since all edges $e \in p$ belong to $\mathtt{supp}(\theta') = \mathtt{supp}(\theta)$. □

4. Analyzing local identifiability

Equipped with the rescaling-invariant embedding $\boldsymbol{\Phi}(\cdot)$ we now establish the claimed local identifiability results. First, we need to introduce notations for the activation status of neurons and paths and use them to provide several expressions of the realization $\boldsymbol{R}_\theta$ before providing the main result of the section, Theorem 5.

4.1. Activation status of neurons and paths, and activation spaces.

The forthcoming analysis heavily involves the activation status of each hidden neuron $\nu \in H$, $a_\nu(\theta, x) = \mathtt{1}_{z_\nu(\theta,x)>0} \in \{0,1\}$, which gives rise to the activation status of each hidden layer $\boldsymbol{a}_\ell(\theta, x) = (a_\nu(\theta, x))_{\nu\in N_\ell}$, $1 \leq \ell \leq L-1$, and the global activation status $\boldsymbol{a}(\theta, x) = (a_\nu(\theta, x))_{\nu\in H} = (\boldsymbol{a}_\ell(\theta, x))_{1\leq\ell\leq L-1}$.

Definition 13. *The activation of a path p (full or partial) is defined as*

$$\alpha_p(\theta, x) := \Pi_{\nu\in H\cap p} a_\nu(\theta, x) \in \{0,1\}$$

where for $p = (p_\ell, \ldots, p_L) \in \mathcal{P}_\ell$ we denote the set of hidden *neurons visited by the path p using the shorthand $H \cap p := \{\nu \in H, \exists i \in [\![\max(\ell, 1), L-1]\!], \nu = p_i\} \subset H$.*

Remark 4. *By convention, a product over an empty set is 1. If p contains no hidden neuron (e.g. , if $p = (\eta) \in \mathcal{P}_L$, $L \geq 1$) its activation is $\alpha_p(\theta, x) = 1$ for every x.*

With $\mathcal{Q} := \cup_{\ell=1}^{L-1}\mathcal{Q}_\ell$ the set of all "partial" paths $q \in (q_\ell, \ldots, q_{L-1})$ from a hidden layer $1 \leq \ell \leq L-1$ to the penultimate layer $L-1$, we define the binary-vector-valued function $\boldsymbol{\alpha}(\theta, x) := (\alpha_q(\theta, x))_{q\in\mathcal{Q}} \in \{0,1\}^{\mathcal{Q}}$. We also define variants that are notably useful to account for output biases

$$\bar{\boldsymbol{a}}_\ell(\theta, x) = \begin{pmatrix}\boldsymbol{a}_\ell(\theta, x)\\ 1\end{pmatrix} \in \{0,1\}^{N_\ell+1} \text{ and } \bar{\boldsymbol{\alpha}}(\theta, x) := \begin{pmatrix}\boldsymbol{\alpha}(\theta, x)\\ 1\end{pmatrix} \in \{0,1\}^{\mathcal{Q}+1}$$

where for any set A, B we use the shorthand $A^{B+1} = A^B \times A$.

To state the connections between non-degeneracy and local S-identifiability from finite sets, it is convenient to observe that the linear space $\mathtt{V}(\theta)$ from Definition 11 can be characterized using simpler linear spaces called *activation spaces*.

Definition 14 (Activation spaces, activation dimension)**.** *The activation spaces associated to* $\theta \in \mathbb{R}^{E\cup\bar{H}}$ *are*

$$\bar{\mathtt{A}}(\theta) := \mathtt{span}\left\{\bar{\boldsymbol{\alpha}}(\theta, x), x \in \mathcal{X}_\theta\right\} \subseteq \mathbb{R}^{\mathcal{Q}+1}. \tag{21}$$

$$\mathtt{A}(\theta) = \mathtt{span}\left\{\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x), x \in \mathcal{X}_\theta\right\} = \boldsymbol{Q}\bar{\mathtt{A}}(\theta) \subseteq \mathbb{R}^{\mathcal{Q}_1} \tag{22}$$

with $\boldsymbol{Q}$ *as in Lemma 10. We define its* activation dimension *as* $\mathtt{actdim}(\theta) = \mathtt{dim}\left(\bar{\mathtt{A}}(\theta)\right)$.

Remark 5. *Observe that if* θ *and* $\tilde{\theta}$ *share the same* $L-1$ *first affine layers* $(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell) = (\widetilde{\boldsymbol{W}}_\ell, \widetilde{b}_\ell)_{\ell=0}^{L-1}$ *then their activation spaces are identical. This holds* even if the dimension of the output layers of $\tilde{\theta}$ and θ differ*.*

Lemma 8. *Viewing* $\mathbb{R}^{\mathcal{P}}$ *as the product of* $N_L \times N_0$ *copies of* $\mathbb{R}^{\mathcal{Q}_1}$ *and* N_L *copies of* $\mathbb{R}^{\mathcal{Q}+1}$*,* $\mathtt{V}(\theta) \subset \mathbb{R}^{\mathcal{P}}$ *is the product of* $N_L \times N_0$ *copies of* $\mathtt{A}^{\perp}(\theta) \subseteq \mathbb{R}^{\mathcal{Q}_1}$ *and* N_L *copies of* $\bar{\mathtt{A}}^{\perp}(\theta)$*.*

The proof is postponed to after Corollary 3 as it uses notations introduced there.

Corollary 2. $\mathtt{V}(\theta) = \{0\}$ *if, and only if,* $\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}+1}$*.*

Proof. If $\mathtt{V}(\theta) = \{0\}$ then by Lemma 8 we have $\bar{\mathtt{A}}^{\perp}(\theta) = \{0\}$ hence $\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}+1}$. Vice-versa if $\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}+1}$ then $\mathtt{A}(\theta) = \boldsymbol{Q}\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}_1}$. By Lemma 8 it follows that $\mathtt{V}(\theta) = \{0\}$, □

4.2. **"Algebraic" expressions of the realization.** We can express the realization using weight matrices, bias vectors and layerwise binary activation vectors. A similar formula is stated without taking the biases into account in [16][Lemma A.2] whereas [35] performs analogous computations for gradient computations, still without biases.

Lemma 9. *Consider* θ *a network parameter of depth* $L \geq 1$*. Denote* $\boldsymbol{I}_0 = \boldsymbol{Id}_{\mathbb{R}^{N_0}}$ *and for each* x *and* $1 \leq \ell \leq L-1$*,* $\boldsymbol{I}_\ell = \mathtt{diag}(\boldsymbol{a}_\ell(\theta, x))$*. The realization of* θ *satisfies*

$$\boldsymbol{R}_\theta(x) = \left(\Pi_{\ell=1}^{L} \boldsymbol{W}_{\boldsymbol{\ell}} \boldsymbol{I}_{\ell-1}\right) x + \sum_{\ell'=1}^{L} \left(\Pi_{\ell=\ell'+1}^{L} \boldsymbol{W}_\ell \boldsymbol{I}_{\ell-1}\right) \boldsymbol{b}_{\ell'} \tag{23}$$

with the convention that a product over an empty set is the identity matrix.

The proof is in Appendix E. To conduct an analysis of the local S-identifiability of a parameter, another expression of $\boldsymbol{R}_\theta$ where the embedding $\boldsymbol{\Phi}(\theta)$ appears more explicitly will be useful. We rewrite (23) using $\boldsymbol{\Phi}(\theta)$ and the activation vector $\bar{\boldsymbol{\alpha}}(\theta, x)$.

Lemma 10. *Consider* θ *a network parameter of depth* $L \geq 2$*. For each* $\eta \in N_L$*, denote*[7]

$$\boldsymbol{\Phi}_\eta^{\mathtt{i}}(\theta) := \left(\Phi_{\mu\to q\to\eta}(\theta)\right)_{q\in\mathcal{Q}_1, \mu\in N_0} \in \mathbb{R}^{\mathcal{Q}_1\times N_0} \tag{24}$$

$$\boldsymbol{\Phi}_\eta^{\mathtt{h}}(\theta) := \begin{pmatrix} \left(\Phi_{q\to\eta}(\theta)\right)_{q\in\mathcal{Q}} \\ \theta_\eta \end{pmatrix} \in \mathbb{R}^{\mathcal{Q}+1} \tag{25}$$

[7]Superscripts $\mathtt{i}$ and $\mathtt{h}$ stand for "input" and "hidden", as $\boldsymbol{\Phi}^{\mathtt{i}}$ is associated to full paths starting from the input layer, while $\boldsymbol{\Phi}^{\mathtt{h}}$ corresponds to partial paths starting from a hidden (or the output) layer.

Up to reshaping, $\boldsymbol{\Phi}(\theta) \in \mathbb{R}^{\mathcal{P}}$ is the concatenation of matrices $\boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta) \in \mathbb{R}^{\mathcal{Q}_1 \times N_0}$ and vectors $\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta) \in \mathbb{R}^{\mathcal{Q}+1}$ over all output neurons $\eta \in N_L$. For each output neuron $\eta \in N_L$ we have

$$\boldsymbol{R}_\theta(x)_\eta = \langle \boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x), \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)x\rangle + \langle \bar{\boldsymbol{\alpha}}(\theta, x), \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta)\rangle \tag{26}$$

where $\boldsymbol{Q} : \mathbb{R}^{\mathcal{Q}+1} \to \mathbb{R}^{\mathcal{Q}_1}$ is the canonical restriction to $\mathcal{Q}_1 \subset \mathcal{Q}$.

The proof of Lemma 10 is in Appendix E. It yields an expression of $\boldsymbol{R}_\theta$ that perfectly fits the upcoming analysis of local S-identifiability. A more abstract (but probably somewhat more digestible) version of the same result implies Property (8) as claimed.

Corollary 3. *Consider θ a network parameter of depth $L \geq 2$. For each $x \in \mathbb{R}^{N_0}$ let $\boldsymbol{L}_{\theta,x}$ be the linear form on $\mathbb{R}^{\mathcal{Q}_1 \times N_0} \times \mathbb{R}^{\mathcal{Q}+1}$ defined as*

$$\boldsymbol{L}_{\theta,x}\{(\boldsymbol{M}, \boldsymbol{v})\} := \langle \boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x), \boldsymbol{M}x\rangle + \langle \bar{\boldsymbol{\alpha}}(\theta, x), \boldsymbol{v}\rangle, \quad \boldsymbol{M} \in \mathbb{R}^{\mathcal{Q}_1 \times N_0}, \quad \boldsymbol{v} \in \mathbb{R}^{\mathcal{Q}+1}$$

Define $\boldsymbol{C}_{\theta,x} \in \mathbb{R}^{N_L \times \mathcal{P}}$ the matrix associated to the linear operator mapping each $\boldsymbol{\phi} \in \mathbb{R}^{\mathcal{P}}$, seen as a reshaped concatenation of matrices $\boldsymbol{\phi}^{\mathtt{i}}_\eta \in \mathbb{R}^{\mathcal{Q}_1 \times N_0}$ and vectors $\boldsymbol{\phi}^{\mathtt{h}}_\eta \in \mathbb{R}^{\mathcal{Q}+1}$ as in Lemma 10, to $\boldsymbol{r} := (r_\eta)_{\eta \in N_L}$, with $r_\eta = \boldsymbol{L}_{\theta,x}\{(\boldsymbol{\phi}^{\mathtt{i}}_\eta, \boldsymbol{\phi}^{\mathtt{h}}_\eta)\}$. We have

$$\boldsymbol{R}_\theta(x) = \boldsymbol{C}_{\theta,x} \cdot \boldsymbol{\Phi}(\theta) \tag{27}$$

We are now equipped with the notations needed to prove Lemma 8. The proof also relies on the following alternative characterization of the set $\mathcal{X}_\theta$ from Definition 10 that we did not find elsewhere. It is proved in Appendix F.

Lemma 11. *Given a parameter θ, consider the open set of input variables x such that $(\theta', z) \mapsto \boldsymbol{a}(\theta', z)$ is locally constant in some neighborhood of (θ, x).*

$$\mathcal{X}'_\theta := \{x \in \mathbb{R}^{N_0} : \exists \epsilon, r > 0,\ \forall (\theta, z) \in B(\theta, \epsilon) \times B(x, r),\ \boldsymbol{a}(\theta', z) = \boldsymbol{a}(\theta, x)\}. \tag{28}$$

with the convention that $\mathcal{X}'_\theta = \mathbb{R}^{N_0}$ if the network depth is $L = 1$. This set coincides exactly with the set $\mathcal{X}_\theta$ from Definition 10.

Proof of Lemma 8. Consider a vector $\boldsymbol{\phi} \in \mathbb{R}^{\mathcal{P}}$ and its representation as $\boldsymbol{\phi}^{\mathtt{i}}_\eta \in \mathbb{R}^{\mathcal{Q}_1 \times N_0}$, $\boldsymbol{\phi}^{\mathtt{h}}_\eta \in \mathbb{R}^{\mathcal{Q}}$, $\eta \in N_L$. By definition $\boldsymbol{\phi} \in \mathtt{V}(\theta)$ if, and only if, $\boldsymbol{C}_{\theta,x}\boldsymbol{\phi} = 0,\ \forall x \in \mathcal{X}_\theta$, i.e., for each $\eta \in N_L$ we have

$$\langle \boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x), \boldsymbol{\phi}^{\mathtt{i}}_\eta x\rangle + \langle \bar{\boldsymbol{\alpha}}(\theta, x), \boldsymbol{\phi}^{\mathtt{h}}_\eta\rangle = 0,\ \forall x \in \mathcal{X}_\theta. \tag{29}$$

By Lemma 11, $x' \mapsto \bar{\boldsymbol{\alpha}}(\theta, x')$ is locally constant in the neighborhood of each $x \in \mathcal{X}_\theta$, hence the left-hand-side in (29) is locally affine with respect to x, and (29) is thus equivalent to

$$\begin{cases} [\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x)]^\top \boldsymbol{\phi}^{\mathtt{i}}_\eta & = \boldsymbol{0}_{1 \times N_0} \\ \langle \bar{\boldsymbol{\alpha}}(\theta, x), \boldsymbol{\phi}^{\mathtt{h}}_\eta\rangle & = 0 \end{cases},\ \forall x \in \mathcal{X}_\theta, \tag{30}$$

that is to say each column of $\boldsymbol{\phi}^{\mathtt{i}}_\eta$, is orthogonal to $\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x)$, and $\boldsymbol{\phi}^{\mathtt{h}}_\eta$ is orthogonal to $\bar{\boldsymbol{\alpha}}(\theta, x)$ for every $x \in \mathcal{X}_\theta$. We conclude using the definition of $\mathtt{A}(\theta), \bar{\mathtt{A}}(\theta)$. □

4.3. **Non-degeneracy and local S-identifiability.** We can now state the main result of this section.

Theorem 5. *Consider $\theta \in \Theta \subseteq \mathbb{R}^{E \cup \bar{H}}$. The following are equivalent:*

i) θ is non-degenerate with respect to Θ;
ii) there is a finite $F \subset \mathcal{X}_\theta$ such that θ is locally S-identifiable from F with respect to Θ.
iii) there is a compact $K \subset \mathcal{X}_\theta$ such that θ is locally S-identifiable from K wrt Θ.

When they hold, the finite set F can be chosen such that

$$\texttt{card}(F) \leq (N_0 + 1)\texttt{actdim}(\theta). \tag{31}$$

Remark 6. *We exhibit in Example 4 a PS-identifiable (hence locally S-identifiable) parameter θ that is degenerate, i.e., not locally S-identifiable from any compact $K \subseteq \mathcal{X}_\theta$.*

Proof. **i) $\Rightarrow$ ii)** Consider $\epsilon > 0$ such that θ is ϵ-non-degenerate with respect to Θ. To establish the existence of F such that θ is locally S-identifiable from F with respect to Θ, we use a Lemma which proof is postponed.

Lemma 12. *Consider $\theta \in \mathbb{R}^{E \cup \bar{H}}$.*

a) There exists $\epsilon > 0$ and a set $F \subset \mathcal{X}_\theta$ of cardinality at most $(N_0+1)\texttt{actdim}(\theta)$ such that: for each $\theta' \in B(\theta, \epsilon)$, if $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta$ on F, then $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$.
b) For every compact set $K \subset \mathcal{X}_\theta$, there exists $\epsilon' > 0$ such that: for each $\theta' \in B(\theta, \epsilon')$, if $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$, then $\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(x)$ for all $x \in K$.

Let ϵ_0, F be given by Lemma 12-a and set $\epsilon_1 := \min(\epsilon_0, \epsilon, \eta/2)$ where $\eta := \min_{i \in \texttt{supp}(\theta)} |\theta_i|$. We will show that θ is ϵ_1-locally S-identifiable from F. For this, consider $\theta' \in \Theta \cap B(\theta, \epsilon_1)$ and assume that $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta$ on K. By Lemma 12-a, since $\theta' \in B(\theta, \epsilon_0)$, we have $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$. Since $\theta' \in B(\theta, \epsilon)$ and θ is ϵ-non-degenerate, this implies $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$ hence (recall that, since θ is non-degenerate, it is admissible by definition) by Lemma 6 we have $\texttt{supp}(\theta') = \texttt{supp}(\theta)$. Since $\theta' \in B(\theta, \eta/2)$ we further have $\texttt{sign}(\theta'_i) = \texttt{sign}(\theta_i)$ for every $i \in \texttt{supp}(\theta)$, hence $\texttt{sign}(\theta') = \texttt{sign}(\theta)$. By Theorem 1 we obtain $\theta' \sim_S \theta$.

ii) $\Rightarrow$ iii) Simply observe that a finite set is compact.

iii) $\Rightarrow$ i) Consider $\epsilon > 0$ such that θ is ϵ-locally identifiable from K with respect to Θ. By Lemma 12-b for the compact set K, there is $\epsilon_0 > 0$ such that: for each $\theta' \in B(\theta, \epsilon_0)$, $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta) \Rightarrow (\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(x), \forall x \in K)$. Set $\epsilon_1 := \min(\epsilon, \epsilon_0)$. We will show that θ is ϵ_1-non-degenerate with respect to Θ. Considering $\theta' \in \Theta \cap B(\theta, \epsilon_1)$ such that $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$ we now show that $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$. Since $\theta' \in B(\theta, \epsilon_0)$ and $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$, we have $\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(x)$ for all $x \in K$. Since $\theta' \in \Theta \cap B(\theta, \epsilon)$ and θ is locally S-identifiable from K with respect to Θ this implies $\theta' \sim_S \theta$, hence by Theorem 1 we have $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$. □

Proof of Lemma 12. We begin with some preliminaries. Since $\bar{\mathtt{A}}(\theta) \subseteq \mathbb{R}^{\mathcal{Q}+1}$, it is finite dimensional hence there is a finite set $\mathcal{Z}_\theta \subset \mathcal{X}_\theta$ such that $\texttt{card}(\mathcal{Z}_\theta) = \texttt{actdim}(\theta)$ and

$$\bar{\mathtt{A}}(\theta) = \texttt{span}\left\{\bar{\boldsymbol{\alpha}}(\theta, z), z \in \mathcal{Z}_\theta\right\}. \tag{32}$$

By definition of $\mathcal{X}_\theta$, for each $z \in \mathcal{Z}_\theta$ there exists $\epsilon(z), r(z) > 0$ such that, for every $\theta' \in B(\theta, \epsilon(z))$ and $x \in B(z, r(z))$, we have $\boldsymbol{a}(\theta', x) = \boldsymbol{a}(\theta, z)$, hence $\bar{\boldsymbol{\alpha}}(\theta', x) = \bar{\boldsymbol{\alpha}}(\theta, z)$. Since $\mathcal{Z}_\theta$ is finite,

$$\epsilon := \min_{z \in \mathcal{Z}_\theta} \epsilon(z) > 0.$$

Consider $\theta' \in B(\theta, \epsilon)$, $z \in \mathcal{Z}_\theta$, $x \in B(z, r(z))$. Since $\bar{\boldsymbol{\alpha}}(\theta', x) = \bar{\boldsymbol{\alpha}}(\theta, z)$, we have for each output neuron $\eta \in N_L$

$$\begin{cases} [\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta', x)]^\top \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') & = [\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta', z)]^\top \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') \\ \bar{\boldsymbol{\alpha}}(\theta', x)^\top \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') & = \bar{\boldsymbol{\alpha}}(\theta', z)^\top \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') \end{cases}$$

hence using (26) we get

$$\boldsymbol{R}_{\theta'}(x)_\eta - \boldsymbol{R}_\theta(x)_\eta = [\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, z)]^\top \left(\boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)\right) x + \bar{\boldsymbol{\alpha}}^\top(\theta, z) \left(\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta)\right). \tag{33}$$

Considering $z \in \mathcal{Z}_\theta$, define $F_z := \{x_i\}_{i=0}^{N_0} \subset B(z, r(z)) \subset \mathbb{R}^{N_0}$ where $x_0 = z$ and for $1 \leq i \leq N_0$, $x_i = z + \frac{r(z)}{2}\delta_i$ with δ_i the i-th vector of the canonical basis. Observe that if $\boldsymbol{u} \in \mathbb{R}^{N_0}, b \in \mathbb{R}$ are such that $\boldsymbol{u}^\top x + b = 0$ for every $x \in F_z$, then $\boldsymbol{u} = \boldsymbol{0}$ (since $r(z)\boldsymbol{u}^\top \delta_i = \boldsymbol{u}^\top(x_i - x_0) = \boldsymbol{u}^\top x_i + b - (\boldsymbol{u}^\top x_0 + b) = 0$ for every i), and therefore $b = 0$ too.

a) The finite set $F := \cup_{z \in \mathcal{Z}_\theta} F_z$ satisfies $F \subset \cup_{z \in \mathcal{Z}_\theta} B(z, r(z)) \subset \mathcal{X}_\theta$. Assume that $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta$ on F where $\theta' \in B(\theta, \epsilon)$. By the preliminaries, this implies that the right hand side in (33) is zero for each $\eta \in N_L$, $z \in \mathcal{Z}_\theta$, $x \in F_z$, hence

$$\begin{cases} [\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, z)]^\top \left(\boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)\right) = \boldsymbol{0}_{1 \times N_0} \\ \bar{\boldsymbol{\alpha}}^\top(\theta, z) \left(\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta)\right) = 0. \end{cases}$$

Since this holds for every $\eta \in N_L$, $z \in \mathcal{Z}_\theta$, in light of (32) this establishes that

$$\forall \eta \in N_L, \begin{cases} \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta) & \in \underbrace{\mathtt{A}^\perp(\theta) \times \ldots \times \mathtt{A}^\perp(\theta)}_{N_0 \text{ times}} \\ \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta) & \in \bar{\mathtt{A}}^\perp(\theta) \end{cases} \tag{34}$$

and we conclude using Lemma 8 and the fact that $\mathtt{card}(F) \leq \mathtt{card}(\mathcal{Z}_\theta) \times (N_0 + 1)$.

b) Since $K \subset \mathcal{X}_\theta$, for each $z \in K$ there are $\epsilon(z), r(z) > 0$ such that: for each $\theta' \in B(\theta, \epsilon(z))$, $x \in B(z, r(z))$, $\bar{\boldsymbol{\alpha}}(\theta', x) = \bar{\boldsymbol{\alpha}}(\theta, z)$. Since K is compact and $K \subset \cup_{z \in K} B(z, r(z))$, there is a finite set $\mathcal{Z} \subset K$ such that $K \subset \cup_{z \in \mathcal{Z}} B(z, r(z))$. Denote $\epsilon' := \min_{z \in \mathcal{Z}} \epsilon(z) > 0$. Considering $\theta' \in B(\theta, \epsilon')$ such that $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$, we now show that $\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(x)$ for each $x \in K$. Given $x \in K$, since there is $z \in \mathcal{Z}_\theta$ such that $x \in B(z, r(z))$, we have

$$\bar{\boldsymbol{\alpha}}(\theta', x) = \bar{\boldsymbol{\alpha}}(\theta, z) = \bar{\boldsymbol{\alpha}}(\theta, x). \tag{35}$$

For each $\eta \in N_L$, since by Lemma 8 $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$ is equivalent to (34), we get

$$[\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta', x)]^\top \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') \overset{(35)}{=} [\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x)]^\top \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') \overset{(34)}{=} [\boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta, x)]^\top \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)$$

$$\bar{\boldsymbol{\alpha}}^\top(\theta', x)\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') \overset{(35)}{=} \bar{\boldsymbol{\alpha}}^\top(\theta, x)\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') \overset{(34)}{=} \bar{\boldsymbol{\alpha}}^\top(\theta, x)\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta).$$

Using (26) we conclude that $\boldsymbol{R}_{\theta'}(x)_\eta = \boldsymbol{R}_\theta(x)_\eta$ for all $\eta \in N_L$, i.e., $\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(x)$. □

5. Identifiability for shallow neural networks

In this section we focus on shallow networks, for which the set $\mathcal{Q} = \mathcal{Q}_1$ of paths is in bijection with the set $H = N_1$ of hidden neurons. Identifying these sets the activation vectors also coincide $\boldsymbol{\alpha}(\theta, x) = \boldsymbol{a}_1(\theta, x) \in \mathbb{R}^{\mathcal{Q}} = \mathbb{R}^{N_1} = \mathbb{R}^H$. After giving a complete characterization of the activation space $\bar{\mathtt{A}}(\theta)$ using the notion of twin neurons, we show that the absence of twin neurons implies non-degeneracy (hence local S-identifiability), and that its combination with irreducibility implies PS-identifiability. Finally, we discuss what happens in the presence of twin neurons.

5.1. Activation spaces and twin neurons.

Whenever θ is admissible, each hidden neuron $\nu \in H$ is not dead, i.e. $\boldsymbol{w}_{\bullet\to\nu} \neq 0$ and $\boldsymbol{w}_{\nu\to\bullet} \neq 0$. According to Definition 8, neurons are twins if their extended vectors $(\boldsymbol{w}_{\bullet\to\nu}, b_\nu)$ are colinear. This defines an equivalence relation, and the hidden layer $H = N_1$ can be partitioned into *equivalence classes of twin neurons*, denoted

$$T_c \subset H, 1 \leq c \leq C.$$

Each equivalence class T_c is partitioned into I_c, J_c, where all neurons in I_c are positive twins, all neurons in J_c are positive twins, and $\nu \in I_c, \nu' \in J_c$ are negative twins. By convention I_c is always non-empty, while J_c may be empty if there are no negative twins in T_c. For each class, we can define a *class signature* vector

$$\mathtt{s}_c = \mathtt{1}_{I_c} - \mathtt{1}_{J_c} \in \mathbb{R}^H,$$

which is zero out of T_c, with ± 1 entries on T_c, and has at least one $+1$ entry. When T_c contains both positive and negative twins, $\mathtt{s}_c$ is only defined up to a global sign. An equivalence class is said to be nontrivial if its cardinal is at least two. Equipped with these notions, we prove in Appendix G the following characterization of activation spaces.

Lemma 13. *Consider an admissible parameter θ on a shallow network architecture. Using the notations introduced above, its activation spaces are*

$$\mathtt{A}(\theta) = \mathtt{span}\left\{\mathtt{1}_H, \mathtt{s}_c, 1 \leq c \leq C\right\} \subseteq \mathbb{R}^H \tag{36}$$

$$\bar{\mathtt{A}}(\theta) = \mathtt{span}\left\{(\mathtt{1}_H, 2), (\mathtt{s}_c, 0), 1 \leq c \leq C\right\} \subseteq \mathbb{R}^{H+1}. \tag{37}$$

5.2. Proof of Lemma 5: no twins implies non-degeneracy.

Lemma 5 is a direct consequence of the combination of Theorem 5 with the following two results.

Lemma 14. *On any network architecture, if $\theta \in \mathbb{R}^{E\cup\bar{H}}$ is admissible and $\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}+1}$ then[8] θ is non-degenerate with respect to any $\Theta \subset \mathbb{R}^{E\cup\bar{H}}$ that contains it.*

Proof. Since $\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}+1}$, by Corollary 2 $\mathtt{V}(\theta) = \{0\}$, hence $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$ is equivalent to $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$. Since θ is admissible, this shows that θ is non-degenerate. □

Lemma 15. *Consider a shallow architecture and $\theta \in \mathbb{R}^{E\cup\bar{H}}$. The equality $\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}+1}$ holds if, and only if, there is no twin. When this holds, $\mathtt{actdim}(\theta) = |H| + 1 = |N_1| + 1$.*

[8]The converse does not hold: there are non-degenerate parameters with $\bar{\mathtt{A}}(\theta) \neq \mathbb{R}^{\mathcal{Q}+1}$, see Lemma 16.

Proof. Equivalence classes of twin neurons form a partition of H, hence $|C| \leq |H| = |\mathcal{Q}|$. By Lemma 13, $\bar{\mathtt{A}}(\theta)$ is the span of $|C|+1$ vectors, hence its dimension is at most $|C|+1$. In the presence of twins we get $|C| < |H|$ hence $\bar{\mathtt{A}}(\theta) \neq \mathbb{R}^{\mathcal{Q}+1}$. In the absence of twins, each equivalence class T_c is trivial, i.e. $|T_c| = 1$. We obtain that $|C| = |H|$, that each signature vector $\mathtt{s}_c$ is a distinct canonical vector δ_c, and obtain $\bar{\mathtt{A}}(\theta) = \mathbb{R}^{\mathcal{Q}+1}$ by Lemma 13. □

5.3. **Proof of Theorem 3: irreducibility and no twins implies PS-identifiability.** By Lemma 3, PS-identifiability from a bounded set with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$ implies that θ has no twins, hence by Lemma 4), it is irreducible (hence admissible), and local S-identifiable, by Theorem 2. For shallow networks, we show that conversely, irreducibility and the absence of twins imply PS-identifiability from a bounded set.

Theorem 6. *Consider N_1, N_1' two finite sets of indices, empty or not*[9]*, and integers $d, k \geq 1$. Consider $\boldsymbol{c} \in \mathbb{R}^k$ and for each $\nu \in N_1$, let $\boldsymbol{v}_\nu \in \mathbb{R}^k$, $\boldsymbol{w}_\nu \in \mathbb{R}^d$ and $b_\nu \in \mathbb{R}$. Define*

$$\boldsymbol{\varphi}(x) = \sum_{\nu \in N_1} \boldsymbol{v}_\nu \mathtt{ReLU}(\langle \boldsymbol{w}_\nu, x\rangle + b_\nu) + \boldsymbol{c}, \qquad x \in \mathbb{R}^d.$$

Similarly define $\boldsymbol{\psi}(x)$ with $\boldsymbol{v}_\nu' \in \mathbb{R}^k$, $\boldsymbol{w}_\nu' \in \mathbb{R}^d$, $b_\nu' \in \mathbb{R}$ for $\nu \in N_1'$, and $\boldsymbol{c}' \in \mathbb{R}^k$.

a) Assume that

- *$\{(\boldsymbol{w}_\nu, b_\nu)\}_{\nu \in N_1}$ are pairwise not collinear, and $\boldsymbol{v}_\nu, \boldsymbol{w}_\nu \neq 0$;*
- *$\{(\boldsymbol{w}_\nu', b_\nu')\}_{\nu \in N_1'}$ are pairwise not collinear, and $\boldsymbol{v}_\nu', \boldsymbol{w}_\nu' \neq 0$.*

If $\boldsymbol{\varphi}(x) = \boldsymbol{\psi}(x)$ for every $x \in \mathbb{R}^d$ then $\mathtt{card}(N_1) = \mathtt{card}(N_1')$.

b) Assume that $\{(\boldsymbol{w}_\nu, b_\nu)\}_{i \in N_1}$ are pairwise not collinear, and

$$\sum_{\nu \in T} \boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top \neq 0, \quad \text{for all non-empty } T \subset N_1. \tag{38}$$

There exists a bounded set $\mathcal{X} \subseteq \mathbb{R}^{N_0}$ (which depends on θ) such that: if $N_1' = N_1$ and $\boldsymbol{\varphi}(x) = \boldsymbol{\psi}(x)$ for every $x \in \mathcal{X}$, then[10] *$\boldsymbol{c} = \boldsymbol{c}'$ and there exists a permutation π of N_1 and $\lambda_\nu > 0$, $\nu \in N_1$ such that*

$$\forall \nu \in N_1 : \boldsymbol{v}_{\pi(\nu)}' = \lambda_\nu^{-1} \boldsymbol{v}_\nu;\ \boldsymbol{w}_{\pi(\nu)}' = \lambda_\nu \boldsymbol{w}_\nu \text{ and } b_{\pi(\nu)}' = \lambda_\nu b_\nu. \tag{39}$$

Proof. As a preliminary, consider $\nu \in N_1$ and denote $\mathcal{V}_\nu := \{x \in \mathbb{R}^{N_0} : \langle \boldsymbol{w}_\nu, x\rangle + b_\nu = 0\}$. Since $\boldsymbol{w}_\nu \neq 0$, the set $\mathcal{V}_\nu$ is a hyperplane which matches the set $\Gamma_\nu(\theta)$ from Definition 10 when considering θ such that $\boldsymbol{\varphi} = \boldsymbol{R}_\theta$. As none of the $(\boldsymbol{w}_\nu, b_\nu)$ is collinear to another, the hyperplanes associated to $\nu \neq \nu' \in N_1$ are distinct. As $\boldsymbol{v}_\nu \neq 0$ for every $\nu \in N_1$ and $\boldsymbol{\varphi}$ is continuous and piecewise affine, this function is differentiable exactly on the complement of $\mathcal{T} := \cup_{\nu \in N_1} \mathcal{V}_\nu$, which is a union of $\mathtt{card}(N_1)$ distinct hyperplanes.

a) Similarly, since none of the $(\boldsymbol{w}_\nu', b_\nu')$ is collinear to another and $\boldsymbol{v}_\nu' \neq 0, \boldsymbol{w}_\nu' \neq 0$ for each $\nu \in N_1'$, the function $\boldsymbol{\psi}$ is differentiable exactly on the complement of a union of $\mathtt{card}(N_1')$ distinct hyperplanes, $\mathcal{T}' = \cup_{\nu \in N'} \mathcal{V}_\nu'$, where $\mathcal{V}_\nu' := \{x \in \mathbb{R}^{N_0} : \langle \boldsymbol{w}_\nu', x\rangle + b_\nu' = 0\}$. Note that $\mathcal{T}$ may be empty since N_1 may be empty, and similarly for $\mathcal{T}'$. Since $\boldsymbol{\varphi} = \boldsymbol{\psi}$, we

[9] We use the convention: $\sum_\emptyset = 0$.

[10] Let us emphasize that here no further assumption is made on $\boldsymbol{w}_\nu', b_\nu', \boldsymbol{v}_\nu'$, $\nu \in N_1$.

have $\mathcal{T} = \mathcal{T}'$ hence $\texttt{card}(N_1) = \texttt{card}(N_1')$, otherwise there would exist one point $x \in \mathbb{R}^d$ where one function would be differentiable and the other not.

b) We now assume $N_1' = N_1$, but make no specific assumption on $\boldsymbol{v}_\nu' \in \mathbb{R}^k$, $\boldsymbol{w}_\nu' \in \mathbb{R}^d$, $b_\nu' \in \mathbb{R}$ for $\nu \in N_1$ or on $\boldsymbol{c}' \in \mathbb{R}^k$. By (38) with $T = \{\nu\}$ we have $\boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top \neq 0$ hence, as in the preliminary, $\mathcal{V}_\nu$, $\nu \in N_1$ are pairwise distinct hyperplanes. Consider an arbitrary hidden neuron $\nu \in N_1$. As the hyperplanes $\{\mathcal{V}_\mu\}_{\mu \in N_1}$ are pairwise distinct, there exist $x_\nu \in \mathcal{V}_\nu$ and $\epsilon_\nu > 0$ such that $\Omega_\nu := B(x_\nu, \epsilon_\nu)$ satisfies $\Omega_\nu \cap \mathcal{T} = \Omega_\nu \cap \mathcal{V}_\nu$. We will show that the result holds with $\mathcal{X} := \cup_{\nu \in N_1} \Omega_\nu$, which is easily seen to be bounded.

From now, assume that $\boldsymbol{\psi}(x) = \boldsymbol{\varphi}(x)$ for every $x \in \mathcal{X}$.

For each $\nu \in \widehat{N}_1 := \{\nu \in N_1 : \boldsymbol{v}_\nu' \neq 0, \boldsymbol{w}_\nu' \neq 0\}$, since $\boldsymbol{w}_\nu' \neq 0$, the set $\mathcal{V}_\nu'$ is a hyperplane. Consider the equivalence relation on $\widehat{N}_1$ defined by: $\nu \sim \mu \Leftrightarrow \mathcal{V}_\nu' = \mathcal{V}_\mu'$, and the resulting quotient set $\bar{N}_1 = \widehat{N}_1 / \sim$. For each equivalence class $\bar{\nu} \in \bar{N}_1$, denote $\mathcal{V}_{\bar{\nu}}'$ the common hyperplane associated to every $\nu \in \bar{\nu}$, and set $\bar{\mathcal{T}} = \cup_{\bar{\nu} \in \bar{N}_1} \mathcal{V}_{\bar{\nu}}'$. We will prove below that there exists an injective map $\pi : N_1 \to \bar{N}_1$ such that $\mathcal{V}_\nu = \mathcal{V}_{\pi(\nu)}'$ for every $\nu \in N_1$. This will imply that $\texttt{card}(\bar{N}_1) \geq \texttt{card}(N_1)$, and since $\texttt{card}(\bar{N}_1) \leq \texttt{card}(\widehat{N}_1) \leq \texttt{card}(N_1)$, it will follow that $\widehat{N}_1 = N_1$ (hence $\boldsymbol{v}_\nu' \neq 0$, $\boldsymbol{w}_\nu' \neq 0$ for every $\nu \in N_1$) and that each equivalence class $\bar{\nu}$ is a singleton. In other words, π is indeed a permutation of N_1, and the hyperplanes $\mathcal{V}_{\{\nu\}}'$, $\nu \in N_1$ are pairwise distinct.

To build π, consider a hidden neuron $\nu \in N_1$. For the sake of contradiction, assume that $\mathcal{V}_{\bar{\mu}}' \neq \mathcal{V}_\nu$ for every $\bar{\mu} \in \bar{N}_1$. This implies the existence of $x_\nu' \in \Omega_\nu \cap \mathcal{V}_\nu$ and of $\epsilon_\nu' > 0$ such that $\Omega_\nu' := B(x_\nu', \epsilon_\nu') \subseteq \Omega_\nu$ and $\Omega_\nu' \cap \bar{\mathcal{T}}' = \emptyset$ and $\Omega_\nu' \cap \mathcal{V}_\nu = \mathcal{V}_\nu$. Since $\Omega_\nu' \cap \bar{\mathcal{T}} = \emptyset$, the function $\boldsymbol{\psi}$ is affine linear on Ω_ν', hence it has constant Jacobian on Ω_ν'. Denote $\Omega_\nu^+ := \{x \in \Omega_\nu' : \langle \boldsymbol{w}_\nu, x \rangle + b_\nu > 0\}$, $\Omega_\nu^- := \{x \in \Omega_\nu' : \langle \boldsymbol{w}_\nu, x \rangle + b_\nu < 0\}$, and observe that both sets are non-empty. For any $x \in \Omega_\nu' \backslash \mathcal{V}_\nu = \Omega_\nu^+ \cup \Omega_\nu^-$, the function $\boldsymbol{\varphi}$ is differentiable and its Jacobian is $\boldsymbol{\varphi}'(x) = \boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top H(\langle \boldsymbol{w}_\nu, x \rangle + b_\nu) + \boldsymbol{d}$ where $\boldsymbol{d} \in \mathbb{R}^k$ and

$$H(t) := \begin{cases} 1, & \text{if } t > 0 \\ 0, & \text{otherwise.} \end{cases}$$

For each $x_\nu^+ \in \Omega_\nu^+, x_\nu^- \in \Omega_\nu^-$ we have $H(\langle \boldsymbol{w}_\nu, x_\nu^+ \rangle + b_\nu) - H(\langle \boldsymbol{w}_\nu, x_\nu^- \rangle + b_\nu) = 1$, hence $\boldsymbol{\varphi}'(x_\nu^+) - \boldsymbol{\varphi}'(x_\nu^-) = \boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top$. As $\boldsymbol{\psi} = \boldsymbol{\varphi}$ on $\mathcal{X} \supseteq \Omega_\nu \supseteq \Omega_\nu'$ and $\boldsymbol{\psi}$ has constant Jacobian on Ω_ν', it follows that $\boldsymbol{v}_\nu \boldsymbol{w}_\nu = 0$, which contradicts our assumptions. Hence, there is $\bar{\mu} \in \bar{N}_1$ such that $\mathcal{V}_{\bar{\mu}}' = \mathcal{V}_\nu$. Since the hyperplanes $\{\mathcal{V}_{\bar{\nu}}'\}_{\bar{\nu} \in \bar{N}_1}$ are pairwise disjoint by construction, such a $\bar{\mu}$ is unique and we define $\pi(\nu) := \bar{\mu}$. Since this holds for every $\nu \in N_1$, we can define the map $\pi : N_1 \to \bar{N}_1$ with $\pi(\nu) := \bar{\mu}$. For $\nu \neq \nu'$ we have $\mathcal{V}_{\pi(\nu')}' = \mathcal{V}_{\nu'} \neq \mathcal{V}_\nu = \mathcal{V}_{\pi(\nu)}'$ since the hyperplanes $\{\mathcal{V}_\nu\}_{\nu \in N_1}$ are pairwise distinct. This proves the injectivity of π. As we have seen, this means that indeed π is a permutation of N_1. Without loss of generality, to simplify notations, we assume from now on that π is the identity.

For each $\nu \in N_1$, since $\mathcal{V}_{\{\nu\}}' = \mathcal{V}_{\pi(\nu)}' = \mathcal{V}_\nu$ there is a nonzero $\lambda_\nu \in \mathbb{R}$ such that

$$(\boldsymbol{w}_\nu', b_\nu') = \lambda_\nu (\boldsymbol{w}_\nu, b_\nu).$$

Reasoning as above, with $\Omega_\nu^\pm$ defined using $\Omega'_\nu := \Omega_\nu$, we obtain that

$$\boldsymbol{\varphi}'(x_\nu^+) - \boldsymbol{\varphi}'(x_\nu^-) = \boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top.$$

for each $x_\nu^+ \in \Omega_\nu^+, x_\nu^- \in \Omega_\nu^-$, and that for each $x \in \Omega_\nu \backslash \mathcal{V}_\nu$, the Jacobian of ψ satisfies $\boldsymbol{\psi}'(x) = \boldsymbol{v}'_\nu (\boldsymbol{w}'_\nu)^\top H(\langle \boldsymbol{w}'_\nu, x\rangle + b'_\nu) + \boldsymbol{d}'$ with some $\boldsymbol{d}' \in \mathbb{R}^k$, hence for each $x_\nu^+ \in \Omega_\nu^+, x_\nu^- \in \Omega_\nu^-$

$$\boldsymbol{\psi}'(x_\nu^+) - \boldsymbol{\psi}'(x_\nu^-) = \boldsymbol{v}'_\nu (\boldsymbol{w}'_\nu)^\top \left(H(\langle \boldsymbol{w}'_\nu, x_\nu^+\rangle + b'_\nu) - H(\langle \boldsymbol{w}'_\nu, x_\nu^-\rangle + b'_\nu) \right).$$

Since $(\boldsymbol{w}'_\nu, b'_\nu) = \lambda_\nu (\boldsymbol{w}_\nu, b_\nu)$ and $\mathtt{sign}(\langle \boldsymbol{w}_\nu, x_\nu^\pm\rangle + b_\nu) = \pm 1$, we have

$$H(\langle \boldsymbol{w}'_\nu, x_\nu^+\rangle + b'_\nu) - H(\langle \boldsymbol{w}'_\nu, x_\nu^-\rangle + b'_\nu) = \mathtt{sign}(\lambda_\nu).$$

Moreover, as $\boldsymbol{\psi} = \boldsymbol{\varphi}$ on $\mathcal{X}$, we have $\boldsymbol{\varphi}'(x_\nu^+) - \boldsymbol{\varphi}'(x_\nu^-) = \boldsymbol{\psi}'(x_\nu^+) - \boldsymbol{\psi}'(x_\nu^-)$, hence

$$\boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top = \boldsymbol{v}'_\nu (\boldsymbol{w}'_\nu)^\top \mathtt{sign}(\lambda_\nu).$$

Since $\boldsymbol{w}'_\nu = \lambda_\nu \boldsymbol{w}_\nu$, this simplifies to

$$\boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top = \boldsymbol{v}'_\nu (\boldsymbol{w}'_\nu)^\top \mathtt{sign}(\lambda_\nu) = \boldsymbol{v}'_\nu \boldsymbol{w}_\nu^\top \lambda_\nu \mathtt{sign}(\lambda_\nu) = |\lambda_\nu| \boldsymbol{v}'_\nu \boldsymbol{w}_\nu^\top \tag{40}$$

Hence, $\boldsymbol{v}'_\nu = \boldsymbol{v}_\nu / |\lambda_\nu|$ for each $\nu \in N_1$.

To conclude, it is enough to prove that $\lambda_\nu > 0$ for every $\nu \in N_\nu$. Using (40), we can re-write the equality $\boldsymbol{\varphi}(x) = \boldsymbol{\psi}(x)$ for every x as follows:

$$\sum_{\nu \in N_1} \boldsymbol{v}_\nu \Big[\mathtt{ReLU}(\langle \boldsymbol{w}_\nu, x\rangle + b_\nu) - |\lambda_\nu|^{-1} \mathtt{ReLU}(\underbrace{\langle \boldsymbol{w}'_\nu, x\rangle + b'_\nu}_{\lambda_\nu(\langle \boldsymbol{w}_\nu, x\rangle + b_\nu)}) \Big] + \boldsymbol{c} - \boldsymbol{c}' = 0. \tag{41}$$

Now, we observe that

$$\mathtt{ReLU}(\langle \boldsymbol{w}_\nu, x\rangle + b_\nu) - |\lambda_\nu|^{-1} \mathtt{ReLU}(\lambda_\nu(\langle \boldsymbol{w}_\nu, x\rangle + b_\nu)) = \begin{cases} 0 & \text{if } \mathtt{sign}(\lambda_\nu) = 1 \\ \langle \boldsymbol{w}_\nu, x\rangle + b_\nu & \text{if } \mathtt{sign}(\lambda_\nu) = -1 \end{cases} \tag{42}$$

We now show that $T := \{\nu \in N_1 \mid \mathtt{sign}(\lambda_\nu) = -1\} = \emptyset$. Using (42), we re-write (41) as:

$$\sum_{\nu \in T} \boldsymbol{v}_\nu (\langle \boldsymbol{w}_\nu, x\rangle + b_\nu) + \boldsymbol{c} - \boldsymbol{c}' = 0. \tag{43}$$

Since this is valid for all $x \in \mathbb{R}^{N_0}$ we get $\boldsymbol{c} = \boldsymbol{c}'$ and $\sum_{\nu \in T} \boldsymbol{v}_\nu \boldsymbol{w}_\nu^\top = 0$. In light of (38) the latter implies $T = \emptyset$, hence $\mathtt{sign}(\lambda_\nu) = 1$ for all $\nu \in N_1$. □

5.4. **Local S-identifiability despite the presence of twins.** It is natural to wonder if there exists shallow networks with twins that are nevertheless either non-degenerate, or locally S-identifiable, or PS-identifiable. Positive twins are excluded (for any network depth) by Lemma 3, hence we can focus on the case where there are $K \geq 1$ nontrivial classes of twins, each made of a single pair of (distinct) negative twins (as any equivalence class with at least three twins necessarily contains two positive ones). We detail here the case $K = 1$ and leave to future work a more detailed analysis of what happens for $K \geq 2$.

Lemma 16 (Single pair of negative twins). *Consider a shallow network architecture. If $\theta \in \Theta \subseteq \mathbb{R}^{E \cup \bar{H}}$ is admissible with a single pair of negative twins, $\{\nu_1, \nu_2\} \subseteq H$, then*

$$\mathtt{A}^{\perp}(\theta) = \{0\} \text{ and } \bar{\mathtt{A}}^{\perp}(\theta) = \mathtt{span}\,\{\delta_{\nu_1} + \delta_{\nu_2} - \delta_\star\} \neq \{0\} \tag{44}$$

where $\delta_\nu \in \mathbb{R}^{H+1}, \nu \in H$ is the ν-th canonical eigenvector, and $\delta_\star = (\mathbf{0}_H, 1)$.

Moreover, if at least one of the following conditions holds:

i) $\boldsymbol{w}_{\nu_1 \to \bullet}$, $\boldsymbol{w}_{\nu_2 \to \bullet}$ are linearly independent (which is only possible if $|N_2| \geq 2$); or
ii) Θ is contained in the set of parameters with zero output bias;

then θ is non-degenerate with respect to Θ. Conversely, if

iii) $\boldsymbol{w}_{\nu_1 \to \bullet}$, $\boldsymbol{w}_{\nu_2 \to \bullet}$ are linearly dependent and θ belongs to the interior of Θ,

then θ is degenerate with respect to Θ.

Remark 7. *Inspecting the proof shows that the assumption in iii) that θ is in the interior of Θ can be relaxed to: for small enough ϵ, each parameter $\theta' \in B(\theta, \epsilon)$ differing from θ only in terms of biases belongs to $\Theta \cap B(\theta, \epsilon)$.*

The proof is in Appendix H. We are now equipped to show with an example that non-degeneracy and local-identifiability are distinct concepts.

Example 4 (Absolute value)**.** *Consider a shallow architecture with scalar input and output and two hidden neurons. The absolute value can be written as $|x| = \mathtt{ReLU}(x) + \mathtt{ReLU}(-x) = \boldsymbol{R}_\theta$ where $\theta = (w_{\mu \to \nu_1} = 1, w_{\mu \to \nu_2} = -1, b_{\nu_1} = b_{\nu_2} = 0, w_{\nu_1 \to \eta} = w_{\nu_2 \to \eta} = 1, b_\eta = 0)$ has a single pair of negative twins. This parameter θ satisfies the following properties*

i) it is not PS-identifiable from any bounded set $\mathcal{X} \subset \mathbb{R}$ (by Lemma 3);
ii) it is not locally S-identifiable from any finite $F \subset \mathcal{X}_\theta$ (i.e., it is degenerate, see below);
iii) it is PS-identifiable (hence locally S-identifiable) from *$\mathcal{X} = \mathbb{R}$;*
iv) it is locally S-identifiable from $F \cup \{0\}$ for some finite set $F \subset \mathcal{X}_\theta$;

The last two points are detailed in Appendix I. Let us detail ii) here. Since $|N_2| = 1$, by Lemma 16-iii) we get that θ is degenerate with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$, i.e. not locally S-identifiable from any finite $F \subseteq \mathcal{X}_\theta$. Indeed, if $F \subseteq \mathcal{X}_\theta = \mathbb{R} \backslash \{0\}$ is finite then $F \subset (-\infty, -t] \cup [t, +\infty)$ for some $t > 0$, and `abs` *coincides on F with (see Figure 2-(c))*

$$\mathtt{ReLU}(x - t) + \mathtt{ReLU}(-(x + t)) + t = \begin{cases} -x, & x \leq -t \\ t, & |x| \leq t \\ x, & x \geq t. \end{cases} = \boldsymbol{R}_{\theta'}(x)$$

where θ' has nonzero biases, so that $\theta' \not\sim_{PS} \theta$.

Example 5 (Revisiting the identity function from Example 1)**.** *The identity function from Example 1 is another example with a single pair of twin neurons. With $\Theta = \mathbb{R}^{E \cup \bar{H}}$ the parameter θ_0 is not locally S-identifiable (from $\mathcal{X} = \mathbb{R}$) as already explained in Example 1. With $\Theta = \Theta_0 \subsetneq \mathbb{R}^{E \cup \bar{H}}$ the set of parameters with zero output bias, θ_0 is on the contrary PS-identifiable from $\mathbb{R}$ (see details in Appendix J). It can also be shown that θ_0 is non-degenerate with respect to Θ_0, using arguments similar to those used in Appendix I to prove*

item iv) of Example 4. This illustrates the fact that, in the presence of a pair of negative twins, many things can happen: the parameter can be PS-identifiable and non- degenerate, or not even locally S-identifiable.

5.5. **Discussion of the role of activation spaces.** For shallow irreducible networks, PS-identifiability from a bounded set is equivalent (Theorem 3) to the absence of twin neurons, which corresponds (Lemma 15) to a completeness property of the activation space that reads $\bar{\mathtt{A}}^{\perp}(\theta) = \{0\}$. The property $\bar{\mathtt{A}}^{\perp}(\theta) = \{0\}$ also implies non-degeneracy (Lemma 14), yet a consequence of Lemma 16 is that the converse does not generally hold (and that the weaker assumption $\mathtt{A}^{\perp}(\theta) = \{0\}$ is no longer sufficient to imply non-degeneracy). An exception occurs for scalar-valued shallow networks.

Lemma 17. *Consider a* scalar-valued *shallow architecture ($|N_2| = 1$). If θ belongs to the interior of $\Theta \subseteq \mathbb{R}^{E\cup\bar{H}}$ and is non-degenerate with respect to Θ then $\bar{\mathtt{A}}^{\perp}(\theta) = \{0\}$.*

This exception is a consequence of the following result.

Lemma 18. *Consider a scalar-valued shallow network architecture. If θ is admissible then there is $0 < C < \infty$ such that: for each $\boldsymbol{z} \in \mathbb{R}^{\mathcal{Q}+1}$, there exists $\theta' \in B(\theta, C\|\boldsymbol{z}\|_\infty)$*

$$\boldsymbol{\Phi}^{\mathtt{i}}_{\eta}(\theta') - \boldsymbol{\Phi}^{\mathtt{i}}_{\eta}(\theta) = \mathbf{0}_{\mathcal{Q}_1\times N_0}, \tag{45}$$

$$\boldsymbol{\Phi}^{\mathtt{h}}_{\eta}(\theta') - \boldsymbol{\Phi}^{\mathtt{h}}_{\eta}(\theta) = \boldsymbol{z} \tag{46}$$

where η is the single output neuron constituting the output layer N_L. The parameters θ and θ' differ only in terms of biases.

Proof. Write $\boldsymbol{z} = (\boldsymbol{y}, \gamma)$ with $\boldsymbol{y} \in \mathbb{R}^{\mathcal{Q}} = \mathbb{R}^H$ and $\gamma \in \mathbb{R}$. For each hidden neuron $\nu \in H = N_1$, denote $v_\nu = v_{\nu\to\eta}$ the unique weight from neuron ν to the single output neuron. For each input neuron $\mu \in N_0$, the μ-th column of $\boldsymbol{\Phi}^{\mathtt{i}}_{\eta}(\theta)$ is $\boldsymbol{\Phi}^{\mathtt{i}}_{\mu\to\eta}(\theta) := (w_{\mu\to\nu}v_\nu)_{\nu\in H}$, and $\boldsymbol{\Phi}^{\mathtt{h}}_{\eta}(\theta) = ((b_\nu v_\nu)_{\nu\in H}, b_\eta)^\top$. To prove the result we define θ' with identical weights as θ, $v'_\nu := v_\nu$, $w'_{\mu\to\nu} := w_{\mu\to\nu}$, and set the output bias to $b'_\eta := b_\eta + \gamma$. This implies $w'_{\mu\to\nu}v'_\nu = w_{\mu\to\nu}v_\nu$ for every $\nu \in H, \mu \in N_0$, hence $\boldsymbol{\Phi}^{\mathtt{i}}_{\eta}(\theta') = \boldsymbol{\Phi}^{\mathtt{i}}_{\eta}(\theta)$. We now seek b'_ν such that $b'_\nu v'_\nu = b_\nu v_\nu + y_\nu$, for each $\nu \in H$. Since θ is admissible, $v_\nu \neq 0$ for all $\nu \in H$, hence we can choose $b'_\nu := b_\nu + y_\nu/v_\nu$. We conclude with C_θ driven by $\min_\nu 1/|v_\nu|$. □

Proof of Lemma 17. We prove the contraposition. Assume that θ is admissible, that it belongs to the interior of $\Theta \subset \mathbb{R}^{E\cup\bar{H}}$, and that $\bar{\mathtt{A}}^{\perp}(\theta) \neq \{0\}$. Since θ is in the interior of Θ, there is $\eta > 0$ such that $B(\theta, \eta) \subseteq \Theta$. For each $0 < \epsilon < \eta$ there exists $\boldsymbol{z} \in \bar{\mathtt{A}}^{\perp}(\theta)$ with norm $\|\boldsymbol{z}\|_\infty = \epsilon/C$ where C is the constant from Lemma 18. Since θ is admissible, by Lemma 18 and the characterization of $\mathtt{V}(\theta)$ (Lemma 8) there exists $\theta' \in B(\theta, C\|\boldsymbol{z}\|_\infty) = B(\theta, \epsilon) = \Theta \cap B(\theta, \epsilon)$ such that $\mathbf{0} \neq \boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$. This shows that θ is degenerate with respect to Θ. □

Remark 8. *The assumption that θ is in the interior of Θ can be relaxed to: each parameter $\theta' \in B(\theta, \epsilon)$ differing from θ only in terms of biases belongs to $\Theta \cap B(\theta, \epsilon)$.*

Acknowledgements

The authors are thankful to François Malgouyres for the interesting discussions on invariant embeddings of linear and ReLU networks we had at different stages of advancement of this work and Joachim Bona-Pellissier for his technical comments on early versions of this article. The authors thank Elisa Riccietti for her feedback that helped a lot improve the readability of this paper. The authors are thankful to Hervé Jégou and Benjamin Graham for their continuous support since the genesis of this project a few years ago.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.
[2] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear approximation and (deep) relu networks, 2019.
[3] Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation, 2020.
[4] Héctor J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 1992.
[5] Paul Kainen, Vera Kurková, Vladik Kreinovich, and Ongard Sirisengtaksin. Uniqueness of network parameterizations and faster learning. *Preprint*, 1994.
[6] Věra Kůrková and Paul C. Kainen. Functionally equivalent feedforward neural networks. *Neural Comput.*, 1993.
[7] Francesca Albertini, Eduardo D. Sontag, and Vincent Maillot. Uniqueness of weights for neural networks. In *Artificial Neural Networks with Applications in Speech and Vision*, 1993.
[8] Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 1994.
[9] David Rolnick and Konrad P. Kording. Reverse-engineering deep relu networks, 2019.
[10] Massimo Fornasier, Timo Klock, and Michael Rauchensteiner. Robust and resource efficient identification of two hidden layer neural networks, 2019.
[11] Mary Phuong and Christoph H Lampert. Functional vs. parametric equivalence of ReLU networks. *ICLR*, 2020.
[12] Francois Malgouyres and Joseph Landsberg. Multilinear compressive sensing and an application to convolutional linear networks. *SIAM*, 2018.
[13] Francois Malgouyres. On the stable recovery of deep structured linear networks under sparsity constraints. *Proceedings of Machine Learning Research*, 2020.
[14] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-Based Capacity Control in Neural Networks. *Journal of Machine Learning Research*, 2015.
[15] Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD - Path-Normalized Optimization in Deep Neural Networks. *NIPS*, 2015.

[16] Qi Meng, Shuxin Zheng, Huishuai Zhang, Wei Chen 0034, Qiwei Ye, Zhi-Ming Ma, Nenghai Yu, and Tie-Yan Liu. G-SGD - Optimizing ReLU Neural Networks in its Positively Scale-Invariant Space. *ICLR*, 2019.

[17] Mingyang Yi, Qi Meng, Wei Chen, Zhi-ming Ma, and Tie-Yan Liu. Positively Scale-Invariant Flatness of ReLU Neural Networks. *arXiv:1903.02237 [cs, stat]*, March 2019. arXiv: 1903.02237.

[18] Pierre Stock, Benjamin Graham, Remi Gribonval, and Hervé Jégou. Equi-normalization of Neural Networks. In *ICLR 2019 - Seventh International Conference on Learning Representations*, pages 1–20, New Orleans, United States, May 2019.

[19] Qunyong Yuan and Nanfeng Xiao. Scaling-Based Weight Normalization for Deep Neural Networks. *IEEE Access*, 7:7286–7295, January 2019. Publisher: IEEE.

[20] Eldad Meller, Alexander Finkelstein, Uri Almog, and Mark Grobman. Same, Same But Different - Recovering Neural Network Quantization Error Through Weight Factorization. *arXiv:1902.01917 [cs, stat]*, February 2019. arXiv: 1902.01917.

[21] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-Free Quantization Through Weight Equalization and Bias Correction. *arXiv:1906.04721 [cs, stat]*, November 2019. arXiv: 1906.04721.

[22] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. Cryptanalytic extraction of neural network models, 2020.

[23] Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *arXiv preprint arXiv:1506.02617*, 2015.

[24] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction, 2019.

[25] Eldad Meller, Alexander Finkelstein, Uri Almog, and Mark Grobman. Same, same but different - recovering neural network quantization error through weight factorization, 2019.

[26] Mingyang Yi, Qi Meng, Wei Chen, Zhi ming Ma, and Tie-Yan Liu. Positively scale-invariant flatness of ReLU neural networks, 2019.

[27] Qi Meng, Shuxin Zheng, Huishuai Zhang, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. $\mathcal{G}$-sgd: Optimizing relu neural networks in its positively scale-invariant space, 2018.

[28] Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns, 2019.

[29] Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations, 2013.

[30] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2014.

[31] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[32] Mary Phuong and Christoph H Lampert. Functional vs. parametric equivalence of relu networks. In *International Conference on Learning Representations*, 2019.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.
[34] Jeremy Bernstein, Jiawei Zhao, Markus Meister, Ming-Yu Liu, Anima Anandkumar, and Yisong Yue. Learning compositional functions via multiplicative weight updates. *Advances in neural information processing systems*, 33, 2020.
[35] David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question?, 2018.

Appendix A. Proof of Lemma 7

To lighten notations we omit the dependence of W, V and $\boldsymbol{S}$ on θ. The proof follows three steps.
1) $\boldsymbol{S}$ is well-defined. Consider $\alpha \in W$ and $\nu \in H$. First, since θ is admissible, there exists at least one path p connecting ν to some output neuron η through edges $e \in E \cap \mathtt{supp}(\theta)$. We wish to show that if p and p' are two such partial paths then $\sum_{e\in p} \alpha_e = \sum_{e\in p'} \alpha_e$. Since θ is admissible, there exists a partial path $\bar{p}$ going from some input neuron μ to ν through edges $e \in E \cap \mathtt{supp}(\theta)$. Since $\bar{p}_\ell = p_\ell = p'_\ell$, define by concatenation the full paths $q = (\bar{p}_0, \dots, \bar{p}_\ell, p_{\ell+1}, \dots, p_L)$ and $q' = (\bar{p}_0, \dots, \bar{p}_\ell, p'_{\ell+1}, \dots, p'_L)$. As q and q' have all their edges in $\mathtt{supp}(\theta)$, we have $\Phi_q(\theta) \neq 0 \neq \Phi_{q'}(\theta)$. Since $\alpha \in W$ and $q, q' \in \mathcal{P}_0$, we have $\Phi_q(\theta) \cdot (\boldsymbol{P}\alpha)_q = \Phi_{q'}(\theta) \cdot (\boldsymbol{P}\alpha)_{q'} = 0$, hence $(\boldsymbol{P}\alpha)_q = (\boldsymbol{P}\alpha)_{q'} = 0$ and

$$\sum_{e\in\overline{p}} \alpha_e + \sum_{e\in p} \alpha_e = \sum_{e\in q} \alpha_e = (\boldsymbol{P}\alpha)_q = 0 = (\boldsymbol{P}\alpha)_{q'} = \sum_{e\in\overline{p}} \alpha_e + \sum_{e\in p'} \alpha_e.$$

Thus, $\sum_{e\in q} \alpha_e = \sum_{e\in q'} \alpha_e$ and $\boldsymbol{S}$ is well-defined.
2) $\boldsymbol{S}$ is injective on V. Note that $\boldsymbol{S}$ is linear hence it is sufficient to show that its kernel is reduced to zero. Let $\alpha \in V$ such that $\boldsymbol{S}\alpha = 0$. Since $\alpha_{\bar{H}} = 0$ and $\mathtt{supp}(\alpha) \subseteq \mathtt{supp}(\theta)$, we have $\alpha_e = 0$ for each $e \in E \backslash \mathtt{supp}(\theta)$, hence it is sufficient to show $\alpha_e = 0$ for any edge $e = \mu \to \nu \in E \cap \mathtt{supp}(\theta)$. Since θ is admissible, there is a partial path $\overline{p}$ going from ν to some output neuron η through edges $e' \in E \cap \mathtt{supp}(\theta)$. Since $e \in E \cap \mathtt{supp}(\theta)$, the extended path $p := \mu \to \overline{p}$ also has all its edges in $E \cap \mathtt{supp}(\theta)$ and joins μ to an output neuron. We distinguish three cases: in the first case, μ, ν are two hidden neurons, and

$$\alpha_e = \sum_{e'\in p} \alpha_{e'} - \sum_{e'\in\overline{p}} \alpha_{e'} = -(\boldsymbol{S}\alpha)_\mu + (\boldsymbol{S}\alpha)_\nu = 0.$$

In the second case, $\mu \in N_0$ is an input neuron, hence p is a full path with edges $e' \in E \cap \mathtt{supp}(\theta)$, so that $\Phi_p(\theta) \neq 0$. Since $\alpha \in W$ it follows that $\sum_{e'\in p} \alpha_{e'} = (\boldsymbol{P}\alpha)_p = 0$ and we also obtain $\alpha_e = 0$. Finally, in the third case, $\nu \in N_L$ is an output neuron hence $\bar{p} = (\nu)$ contains no edge, so that $\sum_{e'\in\bar{p}} \alpha_{e'} = 0$ and we get $\alpha_e = 0$ as well.

c) $\boldsymbol{S}$ is surjective. Consider $\beta \in \mathbb{R}^H$, and $\alpha \in \mathbb{R}^{E\cup\bar{H}}$ as defined around (18). Consider $p = (p_\ell, \dots, p_L) \in \mathcal{P}$ (with $0 \le \ell \le L-1$) a full or partial path going from some neuron $\mu = p_\ell \in N_0 \cup H$ to an arbitrary output neuron $\eta = p_L \in N_L$ through edges $e \in E \cap \mathtt{supp}(\theta)$. In the case of a full path, $\mu \in N_0$ and

$$\sum_{e\in p} \alpha_e = -\beta_{p_{L-1}} + \sum_{j=1}^{L-1} (\beta_{p_j} - \beta_{p_{j-1}}) + \beta_{p_0} = 0.$$

As this holds for any full path with edges $e \in \mathtt{supp}(\theta)$, and since $\alpha_{\bar{H}} = 0$, we get $\alpha \in W$. Since $\alpha_e = 0$ for $e \in E\backslash\mathtt{supp}(\theta)$ we have indeed $\mathtt{supp}(\alpha) \subseteq \mathtt{supp}(\theta)$ hence $\alpha \in V$.

In the case of a partial path ($\ell \ge 1$), we have $\mu \in H$ and similarly

$$\sum_{e\in p} \alpha_e = -\beta_{p_{L-1}} + \sum_{j=\ell+1}^{L-1} (\beta_{p_j} - \beta_{p_{j-1}}) = -\beta_{p_\ell} = -\beta_\mu$$

hence $(\boldsymbol{S}\alpha)_\mu = \beta_\mu$. As this holds for any $\mu \in H$, this shows that $(\boldsymbol{S}\alpha) = \beta$.

Appendix B. Proof of Theorem 2

Denote $\eta = \min_{i\in\mathtt{supp}(\theta)} |\theta_i|$. Assume by contradiction that θ is not locally S-identifiable from $\mathcal{X}$ with respect to Θ. This implies that for each $n \ge 1$ there is $\theta_n \in \Theta \cap B(\theta, \min(\eta, 1/n))$ which is not scaling-equivalent to θ such that $\boldsymbol{R}_{\theta_n} = \boldsymbol{R}_\theta$ on $\mathcal{X}$. For $n \ge 1/\epsilon$, since θ is PS-identifiable from $\mathcal{X}$ with respect to Θ and since $\theta_n \in \Theta$ satisfies $\boldsymbol{R}_{\theta_n} = \boldsymbol{R}_\theta$ on $\mathcal{X}$, we have $\theta_n \sim_{PS} \theta$, hence there is a permutation $\pi_n \in \mathfrak{S}_G$ such that $\pi_n \circ \theta_n \sim_S \theta$, hence by Theorem 1

$$\mathtt{sign}(\pi_n \circ \theta_n) = \mathtt{sign}(\theta), \qquad \text{and} \qquad \boldsymbol{\Phi}(\pi_n \circ \theta_n) = \boldsymbol{\Phi}(\theta). \tag{47}$$

Since the set of permutations is finite, there exists $\pi \in \mathfrak{S}_G$, and an increasing subsequence n_k such that $\pi_{n_k} = \pi$ for each $k \ge 1$. By (47), for every k we have

$$\mathtt{sign}(\pi \circ \theta_{n_k}) = \mathtt{sign}(\theta) \qquad \text{and} \qquad \boldsymbol{\Phi}(\pi \circ \theta_{n_k}) = \boldsymbol{\Phi}(\theta). \tag{48}$$

There is a permutation matrix $\boldsymbol{\Pi} \in \mathbb{R}^{\mathcal{P}\times\mathcal{P}}$ such that $\boldsymbol{\Phi}(\pi \circ \theta') = \boldsymbol{\Pi}\boldsymbol{\Phi}(\theta')$ for all $\theta' \in \mathbb{R}^{E\cup\bar{H}}$. Since $\lim_{k\to\infty} \theta_{n_k} = \theta$, by continuity of $\boldsymbol{\Phi}$ we obtain $\boldsymbol{\Pi}\boldsymbol{\Phi}(\theta) = \boldsymbol{\Phi}(\pi \circ \theta) = \boldsymbol{\Phi}(\theta)$ hence for every $k \ge 1$

$$\boldsymbol{\Phi}(\theta) = \boldsymbol{\Pi}^{-1}\boldsymbol{\Phi}(\theta) \overset{(48)}{=} \boldsymbol{\Pi}^{-1}\boldsymbol{\Pi}\boldsymbol{\Phi}(\theta'_{n_k}) = \boldsymbol{\Phi}(\theta_{n_k}). \tag{49}$$

Since θ is admissible, by Corollary 1, the equality $\boldsymbol{\Phi}(\theta_{n_k}) = \boldsymbol{\Phi}(\theta)$ implies $\mathtt{supp}(\theta_{n_k}) = \mathtt{supp}(\theta)$. This implies that $\mathtt{sign}((\theta_{n_k})_i) = 0 = \mathtt{sign}(\theta_i)$ for each $i \notin \mathtt{supp}(\theta)$. Since $\theta_{n_k} \in B(\theta, \eta)$ for each k, we also have $\mathtt{sign}((\theta'_{n_k})_i) = \mathtt{sign}(\theta_i) \in \{-1, 1\}$ for every $i \in \mathtt{supp}(\theta)$, hence $\mathtt{sign}(\theta_{n_k}) = \mathtt{sign}(\theta)$ for every k. Since θ is admissible, by Theorem 1, the fact that $\mathtt{sign}(\theta_{n_k}) = \mathtt{sign}(\theta)$ and $\boldsymbol{\Phi}(\theta_{n_k}) = \boldsymbol{\Phi}(\theta)$ implies $\theta'_n \sim_S \theta$. This contradicts our assumption that $\theta'_n \not\sim_S \theta$ for every n.

Appendix C. Proof of Lemma 3

We will prove the contraposition using the following observation.

Fact 4. *Consider $\alpha, \beta \in \mathbb{R}$ and $M > 0$. For any $t \in [-M, \infty)$ we have*

$$\alpha\texttt{ReLU}(t) + \beta\texttt{ReLU}(-t) = \begin{cases} \alpha t, & t \geq 0 \\ -\beta t, & t \leq 0 \end{cases} = (\alpha+\beta)\texttt{ReLU}(t) - \beta\texttt{ReLU}(t+M) + M.$$

Assuming that θ has twins, consider a hidden layer $1 \leq \ell \leq L-1$ and $T \subseteq N_\ell$ a pair of twin neurons $T = \{\nu_1, \nu_2\}$. Denote $\boldsymbol{w}_i = \boldsymbol{w}_{\bullet\to\nu_i}$, $b_i = b_{\nu_i}$, $\boldsymbol{v}_i = \boldsymbol{w}_{\nu_i\to\bullet}$. As these neurons are twins, there is $\lambda \in \mathbb{R}$ such that for every $x \in \mathbb{R}^{N_0}$,

$$z_{\nu_2}(\theta, x) = \langle \boldsymbol{w}_2, y_{\ell-1}(\theta, x)\rangle + b_2 = \lambda\left(\langle \boldsymbol{w}_1, y_{\ell-1}(\theta, x)\rangle + b_1\right) = \lambda z_{\nu_1}(\theta, x)$$

In the case of positive twins we have $\lambda > 0$, hence $y_{\nu_1}(\theta, x) = \lambda y_{\nu_2}(\theta, x)$ for every x. Given $\epsilon > 0$ consider $\theta'(\epsilon)$ obtained by keeping unchanged all weights and biases in θ except the weights outgoing from neurons ν_i, $i = 1, 2$: $\boldsymbol{v}'_1 = \boldsymbol{v}_1 + \lambda\epsilon \mathtt{1}_{N_{\ell+1}}$, $\boldsymbol{v}'_2 = \boldsymbol{v}_2 - \epsilon \mathtt{1}_{N_{\ell+1}}$. Since the linear layers and biases of hidden neurons up to layer ℓ are unchanged, we have $\boldsymbol{y}_\ell(\theta, x) = \boldsymbol{y}_\ell(\theta', x)$ for all x. For every neuron $\nu \in N_\ell \backslash T$, since the outgoing weights are unchanged, we get for every x

$$y_\nu(\theta', x)\boldsymbol{w}'_{\nu\to\bullet} = y_\nu(\theta, x)\boldsymbol{w}_{\nu\to\bullet}.$$

Moreover, since $y_{\nu_2}(\theta, x) = \lambda y_{\nu_1}(\theta, x)$ and $\boldsymbol{y}_\ell(\theta, x) = \boldsymbol{y}_\ell(\theta', x)$, we obtain for every x

$$\begin{aligned} y_{\nu_1}(\theta', x)\boldsymbol{w}'_{\nu_1\to\bullet} + y_{\nu_2}(\theta', x)\boldsymbol{w}'_{\nu_2\to\bullet} &= y_{\nu_1}(\theta, x)\boldsymbol{v}'_1 + y_{\nu_2}(\theta, x)\boldsymbol{v}'_2 \\ &= y_{\nu_1}(\theta, x)(\boldsymbol{v}_1 + \lambda\epsilon\mathtt{1}_{N_{\ell+1}}) + \lambda y_{\nu_1}(\theta, x)(\boldsymbol{v}_2 - \epsilon\mathtt{1}_{N_{\ell+1}}) \\ &= y_{\nu_1}(\theta, x)\boldsymbol{v}_1 + \lambda y_{\nu_1}(\theta, x)\boldsymbol{v}_2 \\ &= y_{\nu_1}(\theta, x)\boldsymbol{w}_{\nu_1\to\bullet} + y_{\nu_2}(\theta, x)\boldsymbol{w}_{\nu_2\to\bullet}. \end{aligned}$$

Summing over all hidden neurons we obtain $\boldsymbol{z}_{\ell+1}(\theta, x) = \boldsymbol{z}_{\ell+1}(\theta', x)$ for every x, and since all the next affine layers are unchanged, we obtain $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta$. It is not difficult to check that $\theta' = \theta'(\epsilon)$ is not scaling equivalent to θ and can be made arbitrarily close to it. This shows that θ is not locally S-identifiable from $\mathbb{R}^{N_0}$. By contraposition, if θ is locally S-identifiable from $\mathbb{R}^{N_0}$ then it has no positive twins.

In the case of negative twins we have $y_{\nu_1}(\theta, x) = \texttt{ReLU}(t)$ and $y_{\nu_2}(\theta, x) = |\lambda|\texttt{ReLU}(-t)$ with $t = t(x) := z_{\nu_1}(\theta, x) = \langle \boldsymbol{w}_1, y_{\ell-1}(\theta, x)\rangle + b_1$. Since $\mathcal{X}$ is bounded there is some finite $M > 0$ such that $|z_{\nu_1}(\theta, x)| \leq M$ for every $x \in \mathcal{X}$. Consider θ' obtained by keeping all weights and biases unchanged from θ except the incoming and outgoing weights of ν_1, ν_2, their biases, and the biases of the neurons of the next layer, $\eta \in N_{\ell+1}$, which are set as:

- $\boldsymbol{w}'_{\bullet\to\nu_1} = \boldsymbol{w}'_{\bullet\to\nu_2} = \boldsymbol{w}_{\bullet\to\nu_1}$;
- $b'_{\nu_1} = b_{\nu_1}$; $b'_{\nu_2} = b_{\nu_1} + M$;
- $\boldsymbol{w}'_{\nu_1\to\bullet} = \boldsymbol{w}_{\nu_1\to\bullet} + |\lambda|\boldsymbol{w}_{\nu_2\to\bullet}$; $\boldsymbol{w}'_{\nu_2\to\bullet} = -|\lambda|\boldsymbol{w}_{\nu_2\to\bullet}$;
- $b'_\eta = b_\eta + M$

For each $\eta \in N_{\ell+1}$, using Fact 4 for $\alpha = w_{\nu_1\to\eta}$, $\beta = |\lambda| w_{\nu_2\to\eta}$ we obtain

$$\begin{aligned}
y_{\nu_1}(\theta,x)w_{\nu_1\to\eta} + y_{\nu_2}(\theta,x)w_{\nu_2\to\eta} &= \alpha\mathtt{ReLU}(t) + \beta\mathtt{ReLU}(-t)\\
&= (\alpha+\beta)\mathtt{ReLU}(t) - \beta\mathtt{ReLU}(t+M) + M\\
&= (w_{\nu_1\to\eta} + |\lambda| w_{\nu_2\to\eta})\mathtt{ReLU}(t) - |\lambda| w_{\nu_2\to\eta}\mathtt{ReLU}(t+M) + M\\
&= w'_{\nu_1\to\eta}\mathtt{ReLU}(t) + w'_{\nu_2\to\eta}\mathtt{ReLU}(t+M) + M\\
&= w'_{\nu_1\to\eta}\mathtt{ReLU}(z_{\nu_1}(\theta',x)) + w'_{\nu_2\to\eta}\mathtt{ReLU}(z_{\nu_2}(\theta',x)) + M\\
&= w'_{\nu_1\to\eta} y_{\nu_1}(\theta',x) + w'_{\nu_2\to\eta} y_{\nu_2}(\theta',x) + M.
\end{aligned}$$

Reasoning as in the case of positive twins we obtain $\boldsymbol{z}_{\ell+1}(\theta',x) = \boldsymbol{z}_{\ell+1}(\theta,x)$ and eventually $\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(x)$ for every x in the bounded set $\mathcal{X}$. Since the sign of $\boldsymbol{w}_{\nu_2\to\bullet}$ has changed, θ' is not PS-equivalent to θ. This shows that θ is not PS-identifiable from $\mathcal{X}$ with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$.

By contraposition, assuming that θ is PS-identifiable from a bounded set $\mathcal{X}$ with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$, there is no negative twin. Besides, by Theorem 2, such a θ is also locally S-identifiable from $\mathcal{X}$ with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$, hence it is locally S-identifiable from $\mathbb{R}^{N_0}$ with respect to $\Theta = \mathbb{R}^{E\cup\bar{H}}$. By the first part of the lemma, we conclude that θ has no positive twins either. Hence, it has no twins.

Appendix D. Proof of Lemma 4

We will use the following observation.

Fact 5. *for $\chi \in \{0,1\}$ and $e = (-1)^\chi$ we have $\mathtt{ReLU}(t) = \chi t + e\mathtt{ReLU}(et)$ for every $t \in \mathbb{R}$.*

Assume for the sake of contradiction that θ is not irreducible: $\boldsymbol{W}_{\ell+1}\boldsymbol{I}_T\boldsymbol{W}_\ell = 0$ for some non-empty $T \subset N_\ell$ with some $1 \le \ell \le L-1$. Denote θ' a network with the same weights and biases as θ except on layers ℓ and $\ell+1$, where $\boldsymbol{W}'_\ell, \boldsymbol{W}'_{\ell+1}$ and $\boldsymbol{b}'_\ell, \boldsymbol{b}'_{\ell+1}$ will soon be described. By an easy induction we have $\boldsymbol{y}_{\ell'}(\theta',x) = \boldsymbol{y}_{\ell'}(\theta,x)$ for $0 \le \ell' \le \ell-1$.

Defining $\boldsymbol{J}_T = \mathtt{diag}(e_\nu)_{\nu\in N_\ell}$ with $e_\nu = -1$ if $\nu \in T$ and $e_\nu = 1$ otherwise, we obtain from Fact 5 that for every vector $\boldsymbol{z}_\ell \in \mathbb{R}^{N_\ell}$, $\mathtt{ReLU}(\boldsymbol{z}_\ell) = \boldsymbol{I}_T\boldsymbol{z}_\ell + \boldsymbol{J}_T\ \mathtt{ReLU}(\boldsymbol{J}_T\boldsymbol{z}_\ell)$. Define $\boldsymbol{W}'_\ell = \boldsymbol{J}_T\boldsymbol{W}_\ell$, $\boldsymbol{W}'_{\ell+1} = \boldsymbol{W}_{\ell+1}\boldsymbol{J}_T$, $\boldsymbol{b}'_\ell = \boldsymbol{J}_T\boldsymbol{b}_\ell$. For each $x \in \mathbb{R}^{N_0}$, since $\boldsymbol{y}_{\ell-1}(\theta,x) = \boldsymbol{y}_{\ell-1}(\theta',x)$ and $\boldsymbol{J}_T^2 = \boldsymbol{Id}_{\mathbb{R}^{N_\ell}}$ we get using the shorthands $\boldsymbol{z}_i = \boldsymbol{z}_i(\theta,x)$, $\boldsymbol{z}'_i = \boldsymbol{z}_i(\theta',x)$, $i \in \{\ell,\ell+1\}$

$$\begin{aligned}
\boldsymbol{z}_\ell =& \boldsymbol{W}_\ell\boldsymbol{y}_{\ell-1}(\theta,x) + \boldsymbol{b}_\ell = \boldsymbol{J}_T\left(\boldsymbol{J}_T\boldsymbol{W}_\ell\boldsymbol{y}_{\ell-1}(\theta,x) + \boldsymbol{J}_T\boldsymbol{b}_\ell\right) = \boldsymbol{J}_T\boldsymbol{z}'_\ell\\
\boldsymbol{z}_{\ell+1} =& \boldsymbol{W}_{\ell+1}\ \mathtt{ReLU}(\boldsymbol{z}_\ell) + \boldsymbol{b}_{\ell+1} = \boldsymbol{W}_{\ell+1}\left(\boldsymbol{I}_T\boldsymbol{z}_\ell + \boldsymbol{J}_T\ \mathtt{ReLU}(\boldsymbol{J}_T\boldsymbol{z}_\ell)\right) + \boldsymbol{b}_{\ell+1}\\
=& \underbrace{\boldsymbol{W}_{\ell+1}\boldsymbol{I}_T\boldsymbol{W}_\ell}_{=0}\boldsymbol{y}_{\ell-1}(\theta,x) + \boldsymbol{W}_{\ell+1}\boldsymbol{I}_T\boldsymbol{b}_\ell + \boldsymbol{W}'_{\ell+1}\ \mathtt{ReLU}(\boldsymbol{z}'_\ell) + \boldsymbol{b}_{\ell+1}\\
=& \boldsymbol{W}'_{\ell+1}\ \mathtt{ReLU}(\boldsymbol{z}'_\ell) + (\boldsymbol{W}_{\ell+1}\boldsymbol{I}_T\boldsymbol{b}_\ell + \boldsymbol{b}_{\ell+1}).
\end{aligned}$$

Defining $\boldsymbol{b}'_{\ell+1} := \boldsymbol{W}_{\ell+1}\boldsymbol{I}_T\boldsymbol{b}_\ell + \boldsymbol{b}_{\ell+1}$, we get $\boldsymbol{z}_{\ell+1}(\theta,x) = \boldsymbol{z}'_{\ell+1}(\theta',x)$ for all x. Since all other layers of θ and θ' are identical, an easy induction yields $\boldsymbol{R}_\theta = \boldsymbol{R}_{\theta'}$, where $\theta' \in \mathbb{R}^{E\cup\bar{H}} = \Theta$. To conclude, we prove below that θ' is not PS-equivalent to θ: this contradicts the assumption that θ is PS-identifiable and concludes the proof.

For the sake of (yet another) contradiction, assume that $\theta' \sim_{PS}$, so that there exists diagonal matrices $\boldsymbol{\Lambda}_{\ell'} \in \mathbb{R}^{N_{\ell'} \times N_{\ell'}}$ with positive entries and permutation matrices $\boldsymbol{\Pi}_{\ell'} \in \mathbb{R}^{N_{\ell'} \times N_{\ell'}}$, $0 \leq \ell' \leq L$, such that $\boldsymbol{\Lambda}_0 = \boldsymbol{\Pi}_0 = \boldsymbol{I}_{N_0}$, $\boldsymbol{\Lambda}_L = \boldsymbol{\Pi}_L = \boldsymbol{I}_{N_L}$, $\boldsymbol{W}'_{\ell'} = \boldsymbol{\Pi}_{\ell'} \boldsymbol{\Lambda}_{\ell'} \boldsymbol{W}_{\ell'} \boldsymbol{\Lambda}_{\ell'-1}^{-1} \boldsymbol{\Pi}_{\ell'-1}^{-1}$, and $\boldsymbol{b}'_{\ell'} = \boldsymbol{\Pi}_{\ell'} \boldsymbol{\Lambda}_{\ell'} \boldsymbol{b}_{\ell'}$ for every $1 \leq \ell' \leq L$. We show by induction that $\boldsymbol{\Lambda}_{\ell'} = \boldsymbol{\Pi}_{\ell'} = \boldsymbol{I}_{N_{\ell'}}$ for every $0 \leq \ell' < \ell$. This trivially holds for $\ell' = 0$. If it holds for some $\ell' < \ell - 1$ then, as $(\boldsymbol{W}'_{\ell'+1}, \boldsymbol{b}'_{\ell'+1}) = (\boldsymbol{W}_{\ell'+1}, \boldsymbol{b}_{\ell'+1})$ by construction of θ', we have

$$\begin{aligned}(\boldsymbol{W}_{\ell'+1}, \boldsymbol{b}_{\ell'+1}) = (\boldsymbol{W}'_{\ell'+1}, \boldsymbol{b}'_{\ell'+1}) &= (\boldsymbol{\Pi}_{\ell'+1} \boldsymbol{\Lambda}_{\ell'+1} \boldsymbol{W}_{\ell'+1} \boldsymbol{\Lambda}_{\ell'}^{-1} \boldsymbol{\Pi}_{\ell'}^{-1}, \boldsymbol{\Pi}_{\ell'+1} \boldsymbol{\Lambda}_{\ell'+1} \boldsymbol{b}_{\ell'+1}) \\ &= \boldsymbol{\Pi}_{\ell'+1} \boldsymbol{\Lambda}_{\ell'+1} (\boldsymbol{W}_{\ell'+1}, \boldsymbol{b}_{\ell'+1}),\end{aligned}$$

i.e., $(\boldsymbol{w}_{\bullet \to \nu}, b_\nu) = \lambda_{\pi(\nu)} (\boldsymbol{w}_{\bullet \to \pi(\nu)}, b_{\pi(\nu)})$ for every $\nu \in N_{\ell'+1}$, with π the permutation of $N_{\ell'+1}$ associated to $\boldsymbol{\Pi}_{\ell'+1}$ and $\boldsymbol{\Lambda}_{\ell'+1} = \mathtt{diag}(\lambda_\nu)_{\nu \in N_{\ell'+1}}$. Since θ has no twin, it follows that π is the identity and $\lambda_\nu = 1$ for every $\nu \in N_{\ell'+1}$, which concludes the induction. Now, since $(\boldsymbol{W}'_\ell, \boldsymbol{b}'_\ell) = \boldsymbol{J}_T(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)$ by construction of θ', we have

$$\boldsymbol{J}_T(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell) = (\boldsymbol{W}'_\ell, \boldsymbol{b}'_\ell) = (\boldsymbol{\Pi}_\ell \boldsymbol{\Lambda}_\ell \boldsymbol{W}_\ell \boldsymbol{\Lambda}_{\ell-1}^{-1} \boldsymbol{\Pi}_{\ell-1}^{-1}, \boldsymbol{\Pi}_\ell \boldsymbol{\Lambda}_\ell \boldsymbol{b}_\ell) = \boldsymbol{\Pi}_\ell \boldsymbol{\Lambda}_\ell (\boldsymbol{W}_\ell, \boldsymbol{b}_\ell).$$

As a result, for each $\nu \in T \neq \emptyset$ we have $-(\boldsymbol{w}_{\bullet \to \nu}, b_\nu) = \lambda_{\pi(\nu)} (\boldsymbol{w}_{\bullet \to \pi(\nu)}, b_{\pi(\nu)})$ where π is the permutation of N_ℓ associated to $\boldsymbol{\Pi}_\ell$ and $\boldsymbol{\Lambda}_\ell = \mathtt{diag}(\lambda_\nu)_{\nu \in N_\ell}$. However, since θ has no twin, $(\boldsymbol{w}_{\bullet \to \nu}, b_\nu)$ is not collinear to any $(\boldsymbol{w}_{\bullet \to \nu'}, b_{\nu'})$, $\nu' \in N_\ell$, $\nu' \neq \nu$, hence $\pi(\nu) = \nu$. It follows the $-(\boldsymbol{w}_{\bullet \to \nu}, b_\nu) = \lambda_\nu (\boldsymbol{w}_{\bullet \to \nu}, b_\nu)$, and as $\lambda_\nu > 0$ we obtain $(\boldsymbol{w}_{\bullet \to \nu}, b_\nu) = 0$, therefore θ is not admissible. However, by Lemma 2, since θ is PS-identifiable with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$, it is admissible. Hence the desired contradiction.

Appendix E. Proof of Lemma 9 and Lemma 10

Proof of Lemma 9. The proof is by induction on L. For $L = 1$, since $\boldsymbol{I}_0$ is the identity

$$\boldsymbol{R}_\theta(x) = \boldsymbol{z}_1(\theta, x) = \boldsymbol{W}_1 x + \boldsymbol{b}_1 = \boldsymbol{W}_1 \boldsymbol{I}_0 x + \boldsymbol{b}_1.$$

With the convention that a product of matrices over an empty index set is the identity, this establishes (23) for $L = 1$. Now, assuming that (23) holds for every network of depth L, let us prove it for θ of depth $L + 1$. For this, observe that $\boldsymbol{z}_L(\theta, x)$ is the realization of a network $\underline{\theta}$ of depth L made of the first L affine layers of θ, hence by the induction hypothesis we can use (23) to get

$$\boldsymbol{z}_L(\theta, x) = \boldsymbol{R}_{\underline{\theta}}(x) = \left(\Pi_{\ell=1}^L \boldsymbol{W}_\ell \boldsymbol{I}_{\ell-1}\right) x + \sum_{\ell'=1}^{L-1} \left(\Pi_{\ell=\ell'+1}^L \boldsymbol{W}_\ell \boldsymbol{I}_\ell\right) \boldsymbol{b}_{\ell'}$$

Since $\boldsymbol{y}_L(\theta, x) = \boldsymbol{a}_L(\theta, x) \odot \boldsymbol{z}_L(\theta, x) = \boldsymbol{I}_L \boldsymbol{z}_L(\theta, x)$ we get

$$\begin{aligned}\boldsymbol{R}_\theta(x) = \boldsymbol{z}_{L+1}(\theta, x) &= \boldsymbol{W}_{L+1} \boldsymbol{y}_L(\theta, x) + \boldsymbol{b}_{L+1} \\ &= \boldsymbol{W}_{L+1} \boldsymbol{I}_L \left(\left(\Pi_{\ell=1}^L \boldsymbol{W}_\ell \boldsymbol{I}_{\ell-1}\right) x + \sum_{\ell'=1}^{L} \left(\Pi_{\ell=\ell'+1}^L \boldsymbol{W}_\ell \boldsymbol{I}_{\ell-1}\right) \boldsymbol{b}_{\ell'} \right) + \boldsymbol{b}_{L+1}.\end{aligned}$$

To conclude simply observe that

$$\boldsymbol{W}_{L+1}\boldsymbol{I}_L\left(\Pi_{\ell=1}^{L}\boldsymbol{W}_\ell\boldsymbol{I}_{\ell-1}\right)=\left(\Pi_{\ell=1}^{L+1}\boldsymbol{W}_\ell\boldsymbol{I}_{\ell-1}\right),$$

$$\boldsymbol{W}_{L+1}\boldsymbol{I}_L\left(\sum_{\ell'=1}^{L}\left(\Pi_{\ell=\ell'+1}^{L}\boldsymbol{W}_\ell\boldsymbol{I}_{\ell-1}\right)\boldsymbol{b}_{\ell'}\right)=\sum_{\ell'=1}^{L}\left(\Pi_{\ell=\ell'+1}^{L+1}\boldsymbol{W}_\ell\boldsymbol{I}_{\ell-1}\right)\boldsymbol{b}_{\ell'}$$

$$\text{and that with } \ell'=L+1,\quad \boldsymbol{b}_{L+1}=\boldsymbol{b}_{\ell'}=\left(\Pi_{\ell=\ell'+1}^{L+1}\boldsymbol{W}_\ell\boldsymbol{I}_{\ell-1}\right)\boldsymbol{b}_{\ell'}. \qquad \square$$

Proof of Lemma 10. With $\mathcal{P}_H := \cup_{\ell=1}^{L-1}\mathcal{P}_\ell$ the set of all paths from a hidden neuron to an output neuron, for each output neuron $\eta \in N_L$ we prove at the end of this section that

$$\boldsymbol{R}_\theta(x)_\eta = \sum_{p\in\mathcal{P}_0, p_L=\eta}\alpha_p(\theta,x)\Phi_p(\theta)x_{p_0} + \sum_{p\in\mathcal{P}_H, p_L=\eta}\alpha_p(\theta,x)\Phi_p(\theta)+\theta_\eta \tag{50}$$

Any $p \in \mathcal{P}_0$ is uniquely written $p = \mu \to q \to \eta$ with $\mu = p_0$ its input neuron, $\eta \in N_L$ its output neuron, and $q = (p_1, \dots, p_{L-1}) \in \mathcal{Q}_1$ a partial path from the first layer to the penultimate layer, and $\alpha_{\mu\to q\to\eta}(\theta, x) = \alpha_q(\theta, x)$ for every $q \in \mathcal{Q}_1$ and any $\mu \in N_0$,$\eta \in N_L$. Similarly, for $1 \le \ell \le L-1$, every partial path $p \in \mathcal{P}_\ell$ starting from the ℓ-th hidden layer and ending at the output layer can be written as $p = q \to \eta$ where $\eta = p_L \in N_L$ and $q \in \mathcal{Q}_\ell$ starts from the ℓ-th hidden layer and ends at the penultimate layer, and we have $\alpha_{q\to\eta}(\theta, x) = \alpha_q(\theta, x)$ for all θ, x. Therefore, (50) can be rewritten as

$$\begin{aligned}
\boldsymbol{R}_\theta(x)_\eta &= \sum_{q\in\mathcal{Q}_1}\alpha_q(\theta,x)\sum_{\mu\in N_0}\Phi_{\mu\to q\to\eta}(\theta)x_\mu + \sum_{q\in\mathcal{Q}}\alpha_q(\theta,x)\Phi_{q\to\eta}(\theta)+\theta_\eta\\
&= \sum_{q\in\mathcal{Q}_1}\alpha_q(\theta,x)[\boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)x]_q + \sum_{q\in\mathcal{Q}+1}[\bar{\boldsymbol{\alpha}}(\theta,x)]_q[\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta)]_q\\
&= \langle \boldsymbol{Q}\bar{\boldsymbol{\alpha}}(\theta,x), \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)x\rangle + \langle\bar{\boldsymbol{\alpha}}(\theta,x), \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta)\rangle.
\end{aligned}$$

where we used that $\bar{\boldsymbol{\alpha}}(\theta, x) := (\boldsymbol{\alpha}(\theta, x), 1)$ with $\boldsymbol{\alpha}(\theta, x) := (\alpha_q(\theta, x))_{q\in\mathcal{Q}}$, and $\boldsymbol{Q}$ is the canonical restriction from $\mathcal{Q}+1$ to $\mathcal{Q}_1$. $\square$

Proof of Equation (50). We prove the result by induction on the number of layers L.

For $L = 1$ we have $\mathcal{P}_0 = \{(\mu, \eta)\}_{\mu\in N_0,\eta\in N_1}$ and $\mathcal{P}_1 = \{(\eta)\}_{\eta\in N_1}$. Since $H = \emptyset$, $\alpha_p(\theta, x) = 1$ for all $p \in \mathcal{P}$ and $\mathcal{P}_H = \emptyset$. We have $\Phi_p(\theta) = w_{\mu\to\eta}$ for all $p = (\mu, \eta) \in \mathcal{P}_0$ and $\Phi_p(\theta) = b_\eta$ for each $p = (\eta) \in \mathcal{P}_1$. It follows that

$$\sum_{\substack{p\in\mathcal{P}_0\\ p_L=\eta}}\alpha_p(\theta,x)\Phi_p(\theta)x_{p_0} + \sum_{\substack{p\in\mathcal{P}_H\\ p_L=\eta}}\alpha_p(\theta,x)\Phi_p(\theta)+\theta_\eta = \sum_{\mu\in N_0}w_{\mu\to\eta}x_\mu + b_\eta = (\boldsymbol{W}_1x+\boldsymbol{b}_1)_\eta = (\boldsymbol{R}_\theta(x))_\eta.$$

This establishes (50) for $L = 1$. Assume now that (50) holds for networks of depth $L \ge 1$. With θ a network of depth $L+1$, observe that $\boldsymbol{z}_L(\theta, x) = \boldsymbol{R}_{\tilde\theta}(x)$ with $\tilde\theta$ the network made

of the first L affine layers of θ. Using the induction hypothesis, we get, for $\nu \in N_{L-1}$,

$$z_\nu(\theta, x) = \sum_{\substack{\tilde{p}\in\widetilde{\mathcal{P}}_0 \\ \tilde{p}_{L-1}=\nu}} \alpha_{\tilde{p}}(\tilde{\theta}, x)\Phi_{\tilde{p}}(\tilde{\theta})x_{\tilde{p}_0} + \sum_{\substack{\tilde{p}\in\widetilde{\mathcal{P}}_H \\ \tilde{p}_{L-1}=\nu}} \alpha_{\tilde{p}}(\tilde{\theta}, x)\Phi_{\tilde{p}}(\tilde{\theta}) \tag{51}$$

with $\widetilde{\mathcal{P}}_0 = \{(p_0, \dots, p_{L-1}) \mid p \in \mathcal{P}_0\}$, $\widetilde{\mathcal{P}_H} = \{(p_\ell, \dots, p_{L-1}) \mid p = (p_\ell, \dots, p_L) \in \mathcal{P}_H\}$. Since $\texttt{ReLU}\left(z_\nu(\theta, x)\right) = a_\nu(\theta, x) z_\nu(\theta, x)$ we get

$$\begin{aligned}
(\boldsymbol{R}_\theta(x))_\eta &= \sum_{\nu\in N_{L-1}} y_\nu(\theta, x) w_{\nu\to\eta} + b_\eta = \sum_{\nu\in N_{L-1}} \texttt{ReLU}\left(z_\nu(\theta, x)\right) w_{\nu\to\eta} + b_\eta \\
&= \sum_{\nu\in N_{L-1}} a_\nu(\theta, x) z_\nu(\theta, x) w_{\nu\to\eta} + b_\eta \\
&= \sum_{\nu\in N_{L-1}} \sum_{\substack{\tilde{p}\in\widetilde{\mathcal{P}}_0 \\ \tilde{p}_{L-1}=\nu}} a_\nu(\theta, x)\alpha_{\tilde{p}}(\tilde{\theta}, x)\Phi_{\tilde{p}}(\tilde{\theta}) w_{\nu\to\eta} x_{\tilde{p}_0} \\
&+ \sum_{\nu\in N_{L-1}} \sum_{\substack{\tilde{p}\in\widetilde{\mathcal{P}}_H \\ \tilde{p}_{L-1}=\nu}} a_\nu(\theta, x)\alpha_{\tilde{p}}(\tilde{\theta}, x)\Phi_{\tilde{p}}(\tilde{\theta}) w_{\nu\to\eta} \\
&+ b_\eta
\end{aligned}$$

For each path such that $\tilde{p}_{L-1} = \nu \in N_{L-1}$ we have $a_\nu(\theta, x)\alpha_{\tilde{p}}(\tilde{\theta}, x)\Phi_{\tilde{p}}(\tilde{\theta}) w_{\nu\to\eta} = \alpha_{\tilde{p}\to\eta}(\theta, x)\Phi_{\tilde{p}\to\eta}(\theta)$, and $p := \tilde{p} \to \eta$ belongs to $\mathcal{P}_0$ (resp. to $\mathcal{P}_H$) if, and only if, $\tilde{p} \in \widetilde{\mathcal{P}}_0$ (resp. $\tilde{p} \in \widetilde{\mathcal{P}}_H$). Thus,

$$\begin{aligned}
(\boldsymbol{R}_\theta(x))_\eta &= \sum_{\nu\in N_{L-1}} \sum_{\substack{\tilde{p}\in\widetilde{\mathcal{P}}_0 \\ \tilde{p}_{L-1}=\nu}} \alpha_{\tilde{p}\to\eta}(\theta, x)\Phi_{\tilde{p}\to\eta}(\theta) x_{\tilde{p}_0} + \sum_{\nu\in N_{L-1}} \sum_{\substack{\tilde{p}\in\widetilde{\mathcal{P}}_H \\ \tilde{p}_{L-1}=\nu}} \alpha_{\tilde{p}\to\eta}(\theta, x)\Phi_{\tilde{p}\to\eta}(\theta) + b_\eta \\
&= \sum_{\substack{p\in\mathcal{P}_0 \\ p_L=\eta}} \alpha_p(\theta, x)\Phi_p(\theta) x_{p_0} + \sum_{\substack{p\in\mathcal{P}_H \\ p_L=\eta}} \alpha_p(\theta, x)\Phi_p(\theta) + b_\eta. \qquad \square
\end{aligned}$$

Appendix F. Proof of Lemma 11

The result is proved by induction on the network's depth. The case $L = 1$ is trivial with the convention that a union over an empty family is empty. For any depth, since

$$(\cup_{\nu\in H}\Gamma_\nu(\theta))^c = \cap_{\nu\in H}\Gamma_\nu^c(\theta) = \cap_{\ell=1}^{L-1}\left(\cap_{\nu\in N_\ell}\Gamma_\nu^c(\theta)\right) = \cap_{\ell=1}^{L-1}\left(\cup_{\nu\in N_\ell}\Gamma_\nu(\theta)\right)^c$$

the result is equivalent to $\mathcal{X}'_\theta = \cap_{\ell=1}^{L-1}\left(\cup_{\nu\in N_\ell}\Gamma_\nu(\theta)\right)^c$, which is the quantity manipulated in the induction. Assume that the result is valid for all parameters of depth L and consider θ a parameter of depth $L+1 \geq 2$. Denoting $\underline{\theta} = g(\theta)$ its restriction to its first L layers, we will show that $\mathcal{X}'_\theta = \mathcal{X}'_{\underline{\theta}} \cap (\cup_{\nu\in N_L}\Gamma_\nu(\theta))^c$. First we prove $(\mathcal{X}'_{\underline{\theta}})^c \cup (\cup_{\nu\in N_L}\Gamma_\nu) \subset (\mathcal{X}'_\theta)^c$.

- if $x \notin \mathcal{X}'_{\underline{\theta}}$ then (by definition of $\mathcal{X}'_{\underline{\theta}}$) the function $(\underline{\theta}', x') \mapsto \boldsymbol{a}(\underline{\theta}', x')$ is not locally constant around $(\underline{\theta}, x)$ hence there exists $1 \leq \ell \leq L-1$ and $\nu \in N_\ell$ such that $a_\nu(\underline{\theta}', x')$ is not locally constant around $(\underline{\theta}, x)$. Since $\ell \leq L-1$, for every θ', x' we

have $a_\nu(\theta', x') = a_\nu(\underline{\theta}', x')$ with $\underline{\theta}' = g(\theta')$ the restriction of θ' to its first L layers. We obtain that $a_\nu(\theta', x')$ is not locally constant around (θ, x), showing that $x \notin \mathcal{X}'_\theta$.
- If $x \in \cup_{\nu \in N_L} \Gamma_\nu(\theta)$, there exists $\nu \in N_L$ such that $x \in \Gamma_\nu(\theta)$ hence $z_\nu(\theta, x) = 0$, the gradient is well-defined, and $\nabla z_\nu(\theta, x) \neq 0$. This implies that the sign of $z_\nu(\theta, x')$ is not locally constant around x, hence $x' \mapsto a_\nu(\theta, x')$ is not locally constant around x, therefore $(\theta', x') \mapsto \boldsymbol{a}(\theta', x')$ is not locally constant around (θ, x). Thus, $x \notin \mathcal{X}'_\theta$.

This establishes equivalently that $\mathcal{X}'_\theta \subset \mathcal{X}'_{\underline{\theta}} \cap (\cup_{\nu \in N_L} \Gamma_\nu)^c$.

Vice-versa, consider $x \in \mathcal{X}'_{\underline{\theta}} \cap (\cup_{\nu \in N_L} \Gamma_\nu(\theta))^c$. Since $x \in \mathcal{X}'_{\underline{\theta}}$, $\boldsymbol{a}(\underline{\theta}', x')$ is locally constant around $(\underline{\theta}, x)$, hence $(\theta', x') \mapsto \boldsymbol{a}_\ell(g(\theta'), x') = \boldsymbol{a}_\ell(\theta', x')$ is locally constant around (θ, x) for each $1 \leq \ell \leq L-1$. There remains to show that $\boldsymbol{a}_L(\theta', x')$ is locally constant around (θ, x). Indeed, since $x \notin \cup_{\nu \in N_L} \Gamma_\nu$, we have $z_\nu(\theta, x) \neq 0$ for every $\nu \in N_L$. By continuity of $(\theta', x') \mapsto \boldsymbol{z}_L(\theta', x')$, there exists a neighborhood of (θ, x) on which $\mathtt{sign}(z_\nu(\theta', x'))$ is constant for every $\nu \in N_L$, hence $\boldsymbol{a}_L(\theta', x')$ is locally constant around (θ, x). Overall, we get that $(\theta, x) \mapsto \boldsymbol{a}(\theta', x')$ is locally constant around (θ, x), i.e. $x \in \mathcal{X}'_\theta$. This concludes the proof that $\mathcal{X}'_{\underline{\theta}} \cap (\cup_{\nu \in N_L} \Gamma_\nu(\theta))^c \subset \mathcal{X}'_\theta$, hence the equality $\mathcal{X}'_\theta = \mathcal{X}'_{\underline{\theta}} \cap (\cup_{\nu \in N_L} \Gamma_\nu(\theta))^c$.

Appendix G. Proof of Lemma 13

First we prove that $\mathtt{s}_c \in \mathtt{A}(\theta)$ and $(\mathtt{s}_c, 0) \in \bar{\mathtt{A}}(\theta)$ for each c. Since θ is admissible, one can check (cf Definition 8) that two hidden neurons $\nu, \nu' \in H$ of a shallow network are:

- positive twins if, and only if, $a_\nu(\theta, x) = a_{\nu'}(\theta, x)$ for all $x \in \mathcal{X}_\theta$;
- negative twins if, and only if, $a_\nu(\theta, x) = 1 - a_{\nu'}(\theta, x)$ for all $x \in \mathcal{X}_\theta$;

Since we are on a shallow architecture, we identify $\mathcal{Q} = \mathcal{Q}_1$ with H and $\boldsymbol{a}(\theta, x)$ with $\boldsymbol{\alpha}(\theta, x)$. Considering the c-th equivalence class T_c of twins, it follows that for every x there is $\epsilon_c(x) \in \{-1, +1\}$ such that

$$2\boldsymbol{\alpha}_{T_c}(\theta, x) = 2\boldsymbol{a}_{T_c}(\theta, x) = \mathtt{1}_{T_c} + \epsilon_c(x) \cdot \mathtt{s}_c \tag{52}$$

where for any $\boldsymbol{u} \in \mathbb{R}^H$, $\boldsymbol{u}_T \in \mathbb{R}^H$ is its restriction to T (which matches $\boldsymbol{u}$ on its coordinates indexed by T and is zero elsewhere), and $\mathtt{1}_H \in \mathtt{R}^H$ is the vector with all entries equal to one, while $\mathtt{1}_T$ is its restriction to T. To continue we use the following result.

Lemma 19. *Consider a shallow network with parameter θ, and $T \subset H$ an equivalence class of twin neurons. There are $x_T^+, x_T^- \in \mathcal{X}_\theta$ such that*

$$|a_\nu(\theta, x_T^+) - a_\nu(\theta, x_T^-)| = \begin{cases} 1, & \text{if } \nu \in T \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

Proof. For each $\nu \in H$ denote $\mathcal{V}_\nu = \{x \in \mathbb{R}^{N_0} : \langle \boldsymbol{w}_{\bullet \to \nu}, x \rangle + b_\nu = 0\}$. Since θ is admissible, $\boldsymbol{w}_{\bullet \to \nu} \neq 0$ for each $\nu \in H$, hence these linear spaces are hyperplanes. The hyperplanes associated to two neurons coincide if, and only if, these neurons are twins. Choose an arbitrary $\nu \in T$. Since $\mathcal{V}_\nu$ is distinct from each of the (finitely many) $\mathcal{V}_{\nu'}$, $\nu' \notin T$, there exists $x_0 \in \mathcal{V}_\nu$ that belongs to the complement of $\cup_{\nu' \notin T} \mathcal{V}_{\nu'}$. As this complement is open, there exists $\epsilon > 0$ such that $B(x_0, \epsilon \|\boldsymbol{w}_{\bullet \to \nu}\|_2)$ does not intersect any of the hyperplanes $\mathcal{V}_{\nu'}$, $\nu' \notin T$. Since $x_T^\pm := x_0 \pm \boldsymbol{w}_{\bullet \to \nu} \epsilon / 2 \in B(x_0, \epsilon \|\boldsymbol{w}_{\bullet \to \nu}\|_2)$ we obtain: $a_{\nu'}(\theta, x_T^+) = a_{\nu'}(\theta, x_T^-)$

for every $\nu' \notin T$, and $\mathtt{sign}(\langle \boldsymbol{w}_{\bullet\to\nu}, x_T^{\pm}\rangle + b_\nu) = \pm 1$ hence $a_\nu(\theta, x_T^+) = 1 - a_\nu(\theta, x_T^-)$. The latter extends to each $\nu' \in T$ by the twin property, and yields the conclusion. □

By Lemma 19 there are $x_c^+, x_c^- \in \mathcal{X}_\theta$ such that

$$|a_\nu(\theta, x_c^+) - a_\nu(\theta, x_c^-)| = \begin{cases} 1, & \text{if } \nu \in T_c \\ 0, & \text{if } \nu \in H \backslash T_c \end{cases} \tag{54}$$

It follows that $\boldsymbol{\alpha}(\theta, x_c^+) - \boldsymbol{\alpha}(\theta, x_c^-) = \boldsymbol{\alpha}_{T_c}(\theta, x_c^+) - \boldsymbol{\alpha}_{T_c}(\theta, x_c^-) = \pm \mathtt{s}_c$. As a result

$$\mathtt{s}_c = \pm \left(\boldsymbol{\alpha}(\theta, x_c^+) - \boldsymbol{\alpha}(\theta, x_c^-)\right) \in \mathtt{A}(\theta)$$
$$(\mathtt{s}_c, 0) = \pm \left(\bar{\boldsymbol{\alpha}}(\theta, x_c^+) - \bar{\boldsymbol{\alpha}}(\theta, x_c^-)\right) \in \bar{\mathtt{A}}(\theta)$$

as claimed. Using (52) and the partition of H into $T_1, \dots, T_C$ we have for any $x \in \mathcal{X}_\theta$

$$2\boldsymbol{\alpha}(\theta, x) = \sum_c 2\boldsymbol{\alpha}_{T_c}(\theta, x) = \sum_c (\mathtt{1}_{T_c} + \epsilon_c(x) \cdot \mathtt{s}_c) = \mathtt{1}_H + \sum_c \epsilon_c(x)\mathtt{s}_c \tag{55}$$

We obtain

$$\mathtt{1}_H = 2\boldsymbol{\alpha}(\theta, x) - \sum_c \epsilon_c(x)\mathtt{s}_c, \tag{56}$$

and since $\boldsymbol{\alpha}(\theta, x) \in \mathtt{A}(\theta)$ and $\mathtt{s}_c \in \mathtt{A}(\theta)$ for all c, it follows that $\mathtt{1}_H \in \mathtt{A}(\theta)$. This proves $\mathtt{span}\{\mathtt{1}_H, \mathtt{s}_c, 1 \le c \le C\} \subseteq \mathtt{A}(\theta)$. Vice-versa, (55) shows $\boldsymbol{\alpha}(\theta, x) \in \mathtt{span}\{\mathtt{1}_H, \mathtt{s}_c, 1 \le c \le C\}$ for every $x \in \mathcal{X}_\theta$, hence $\mathtt{A}(\theta) \subseteq \mathtt{span}\{\mathtt{1}_H, \mathtt{s}_c, 1 \le c \le C\}$. By (56) we also get

$$(\mathtt{1}_H, 2) = 2(\boldsymbol{\alpha}(\theta, x), 1) - \sum_c \epsilon_c(x)(\mathtt{s}_c, 0),$$

and since $(\boldsymbol{\alpha}(\theta, x), 1) = \bar{\boldsymbol{\alpha}}(\theta, x) \in \bar{\mathtt{A}}(\theta)$ and $(\mathtt{s}_c, 0) \in \bar{\mathtt{A}}(\theta)$, we get $(\mathtt{1}_H, 2) \in \bar{\mathtt{A}}(\theta)$. This proves $\mathtt{span}\{(\mathtt{1}_H, 2), (\mathtt{s}_c, 0), 1 \le c \le C\} \subseteq \bar{\mathtt{A}}(\theta)$, and also implies

$$2\bar{\boldsymbol{\alpha}}(\theta, x) = (\mathtt{1}_H, 2) + \sum_c \epsilon_c(x)(\mathtt{s}_c, 0)$$

hence $\bar{\mathtt{A}}(\theta) \subseteq \mathtt{span}\{(\mathtt{1}_H, 2), (\mathtt{s}_c, 0), 1 \le c \le C\}$.

Appendix H. Proof of Lemma 16

We use the shorthands $\boldsymbol{w}_\nu = \boldsymbol{w}_{\bullet\to\nu}$, $\boldsymbol{v}_\nu = \boldsymbol{w}_{\nu\to\bullet}$.

Given the assumption there are $C = |H| - 1$ classes of twin neurons, all being trivial except one made of a pair of negative twins $\{\nu, \nu'\}$. Without loss of generality we enumerate the neurons and their classes such that $T_1 = \{\nu_1, \nu_2\} = \{\nu, \nu'\}$ and $T_c = \{\nu_{c+1}\}$, $2 \le c \le C = |H| - 1$. First we establish that, with this numbering,

$$\bar{\mathtt{A}}^\perp(\theta) = \mathtt{span}\{(1, 1, 0, \dots, 0, -1)\} \text{ and } \mathtt{A}^\perp(\theta) = \{0\}. \tag{57}$$

The signatures of the classes are $\mathtt{s}_1 = \delta_1 - \delta_2$ and $\mathtt{s}_c = \delta_{c+1}$, $2 \le c \le C$. By Lemma 13 we have $\mathtt{A}(\theta) = \mathtt{span}\{\mathtt{1}_H, \mathtt{s}_c, 1 \le c \le C\}$. It is not difficult to check [11] that the $C + 1 = |H|$

[11] If, instead of a single pair of negative twins, we consider a single pair of *positive* twins, then $\mathtt{s}_1 = \delta_1 + \delta_2$ and the spanning vectors of $\mathtt{A}(\theta)$ become linearly *dependent*, with $\mathtt{A}^\perp(\theta) = \mathtt{span}\{(1, -1, 0, \dots, 0)\} \neq \{0\}$.

spanning vectors are linearly independent, hence $\mathtt{A}(\theta) = \mathbb{R}^H$ and $\mathtt{A}^\perp(\theta) = \{0\}$. Now, consider $\boldsymbol{v} = (v_1, \dots, v_{|H|+1}) \in \bar{\mathtt{A}}^\perp(\theta)$. By Lemma 13, this vector is orthogonal to each $(\mathtt{s}_c, 0)$, $1 \le c \le C$, and to $(\mathtt{1}_H, 2)$. For $2 \le c \le C$, orthogonality to $(\mathtt{s}_c, 0) = (\delta_{c+1}, 0)$ implies $v_{c+1} = 0$, hence $\boldsymbol{v} = (\alpha, \beta, 0, \dots, 0, \gamma)$ for some $\alpha, \beta, \gamma \in \mathbb{R}$. Orthogonality to $(\mathtt{s}_1, 0) = (1, -1, 0, \dots, 0)$ implies $\beta = \alpha$, and orthogonality to $(\mathtt{1}_H, 2)$ implies $\gamma = -\alpha$, hence $\boldsymbol{v}$ is proportional to $(1, 1, 0, \dots, 0, -1)$ as claimed. Since $\bar{\mathtt{A}}(\theta)$ is spanned by $C + 1 = |H|$ vectors, its dimension is at most $|H|$, hence the dimension of $\bar{\mathtt{A}}^\perp(\theta)$ is at least one. This concludes the proof that $\bar{\mathtt{A}}^\perp(\theta) = \mathtt{span}\{(1, 1, 0, \dots, 0, -1)\}$.

Since θ is admissible, there is an input neuron $\mu \in N_0$ such that $w_{\mu\to\nu_1} \neq 0$. Since ν_1,ν_2 are twins, we also have $w_{\mu\to\nu_2} \neq 0$. Let $\epsilon_0 := \min_{1\le j\le 2} |w_{\mu\to\nu_j}|/2$. Consider $\theta' \in B(\theta, \epsilon_0)$ such that $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$. First, observe that $w'_{\mu\to\nu_j} \neq 0$ for $j = 1, 2$. Then, in light of Lemma 8 and (57), for every $\eta \in N_2$, we have $\boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') = \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)$ and

$$\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta) \in \mathtt{span}\{(1, 1, 0, \dots, 0, -1)\}, \tag{58}$$

hence $b'_{\nu_{c+1}} w'_{\nu_{c+1}\to\eta} = b_{\nu_{c+1}} w_{\nu_{c+1}\to\eta}$ for $2 \le c \le C$, and there are scalars $\lambda_\eta \in \mathbb{R}$ such that

$$b'_{\nu_j} w'_{\nu_j\to\eta} - b_{\nu_j} w_{\nu_j\to\eta} = \lambda_\eta,\ \forall j \in \{1, 2\} \qquad \text{and } b'_\eta - b_\eta = -\lambda_\eta. \tag{59}$$

When Θ is the set of parameters with zero output biases, the fact that $\theta, \theta' \in \Theta$ implies $b'_\eta = b_\eta = 0$, hence $\lambda_\eta = 0$ and $\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') = \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta)$ for every $\eta \in N_2$. We show below that the same holds for arbitrary Θ when $\boldsymbol{w}_{\nu_1\to\bullet}$ and $\boldsymbol{w}_{\nu_2\to\bullet}$ are linearly independent. This implies $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$, hence θ is then ϵ-non-degenerate with respect to Θ.

Indeed, the equality $\boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') = \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)$ for all $\eta \in N_2$ implies that for $1 \le j \le 2$,

$$w'_{\mu\to\nu_j} w'_{\nu_j\to\eta} = w_{\mu\to\nu_j} w_{\nu_j\to\eta},\ \forall \eta \in N_2, \tag{60}$$

and since $w'_{\mu\to\nu_j} \neq 0$ for $j = 1, 2$, we obtain from (59) and (60) that for each $\eta \in N_2$,

$$\begin{aligned}\lambda_\eta = b'_{\nu_j} w'_{\nu_j\to\eta} - b_{\nu_j} w_{\nu_j\to\eta} &= b'_{\nu_j} \frac{w'_{\mu\to\nu_j} w'_{\nu_j\to\eta}}{w'_{\mu\to\nu_j}} - b_{\nu_j} w_{\nu_j\to\eta} \\ &= b'_{\nu_j} \frac{w_{\mu\to\nu_j} w_{\nu_j\to\eta}}{w'_{\mu\to\nu_j}} - b_{\nu_j} w_{\nu_j\to\eta} = \left(b'_{\nu_j} \frac{w_{\mu\to\nu_j}}{w'_{\mu\to\nu_j}} - b_{\nu_j} \right) w_{\nu_j\to\eta}\end{aligned}$$

We obtain $\boldsymbol{\lambda} = x_j \boldsymbol{w}_{\nu_j\to\bullet}$, $j = 1, 2$ where $\boldsymbol{\lambda} := (\lambda_\eta)_{\eta\in N_2}$ and $x_j := b'_{\nu_j} \frac{w_{\mu\to\nu_j}}{w'_{\mu\to\nu_j}} - b_{\nu_j}$. Since $\boldsymbol{w}_{\nu_1\to\bullet}$ and $\boldsymbol{w}_{\nu_2\to\bullet}$ are linearly independent, it follows that $x_1 = x_2 = 0$, hence $\boldsymbol{\lambda} = \mathbf{0}$.

Assume now that $\boldsymbol{w}_{\nu_1\to\bullet}$ and $\boldsymbol{w}_{\nu_2\to\bullet}$ are linearly *dependent*, and recall that since θ is admissible they are both nonzero vectors, hence $\boldsymbol{w}_{\nu_2\to\bullet} = \alpha \boldsymbol{w}_{\nu_2\to\bullet}$ for some $\alpha \neq 0$. Consider

$0 < \epsilon < \epsilon_0$ and set θ' as follows:

$$\begin{aligned}
\boldsymbol{W}'_\ell &= \boldsymbol{W}_\ell,\ 1 \leq \ell \leq 2;\\
b'_\nu &= b_\nu,\ \nu \in H \backslash \{\nu_1, \nu_2\};\\
b'_{\nu_1} &= b_{\nu_1} + \gamma\epsilon;\\
b'_{\nu_2} &= b_{\nu_2} + \gamma\epsilon/\alpha, j = 1, 2;\\
b'_\eta &= b_\eta - w_{\nu_1 \to \eta}\gamma\epsilon, \eta \in N_2,
\end{aligned}$$

with $0 < \gamma < \min(1, |\alpha|, 1/\|\boldsymbol{w}_{\nu_1 \to \bullet}\|_\infty)$ so that $\theta' \in B(\theta, \epsilon)$. Since the weights of θ' and θ coincide we have $\boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta') = \boldsymbol{\Phi}^{\mathtt{i}}_\eta(\theta)$ for every $\eta \in N_2$. It is not difficult to check that, with $\lambda_\eta := w_{\nu_1 \to \eta}\epsilon$, we also have $\boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta') - \boldsymbol{\Phi}^{\mathtt{h}}_\eta(\theta) = \lambda_\eta(1, 1, 0, \ldots, 0, -1)$, hence $\boldsymbol{\Phi}(\theta') - \boldsymbol{\Phi}(\theta) \in \mathtt{V}(\theta)$. Yet, $\boldsymbol{\Phi}(\theta') \neq \boldsymbol{\Phi}(\theta)$ since $\boldsymbol{\lambda} := (\lambda_\eta)_{\eta \in N_2} = \epsilon \boldsymbol{w}_{\nu_1 \to \bullet} \neq \mathbf{0}$. Assuming that θ belongs to the interior of Θ, we have $\theta' \in \Theta \cap B(\theta, \epsilon)$ for small enough ϵ. It follows that θ is degenerate.

Appendix I. Details on Example 4

θ is PS-identifiable from $\mathcal{X} = \mathbb{R}$. Consider an arbitrary $\theta' \in \Theta = \mathbb{R}^{E \cup \bar{H}}$. If $\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(x) = |x|$ on $\mathbb{R}$ then θ' is admissible (otherwise its realization would be, up to an additive constant, proportional to a single shifted version of the ReLU, which would prevent it from being equal to $\boldsymbol{R}_\theta = \mathtt{abs}$) hence $w'_{\nu \to \nu_i} \neq 0$, $i = 1, 2$. Writing $\alpha_i = |w'_{\mu \to \nu_i}|\ w'_{\nu_i \to \eta}$ and $\beta_i = -b'_{\nu_i}/|w'_{\mu \to \nu_i}|$, and $s_i = \mathtt{sign}(w'_{\mu \to \nu_i}) \in \{-1, +1\}$ for $i = 1, 2$, we have $\alpha_i \neq 0$ and

$$\boldsymbol{R}_{\theta'}(x) = \alpha_1 \mathtt{ReLU}(s_1(x - s_1\beta_1)) + \alpha_2 \mathtt{ReLU}(s_2(x - s_2\beta_2)) + b'_\eta,\ \forall x \in \mathbb{R}.$$

If we had $s_1\beta_1 \neq s_2\beta_2$, $\boldsymbol{R}_{\theta'}$ would be non-differentiable at two distinct points $s_1\beta_1, s_2\beta_2$. However $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta = \mathtt{abs}$ is differentiable on $\mathbb{R}\backslash\{0\}$, hence $s_1\beta_1 = s_2\beta_2$, and a similar reasoning yields $s_1\beta_1 = s_2\beta_2 = 0$. Since $|s_i| = 1$, we get $\beta_1 = \beta_2 = 0$ and

$$\boldsymbol{R}_{\theta'}(x) = \alpha_1 \mathtt{ReLU}(s_1 x) + \alpha_2 \mathtt{ReLU}(s_2 x) + b'_\eta,\ \forall x \in \mathbb{R}.$$

If we had $s_1 = s_2$, the realization would be $(\alpha_1 + \alpha_2)\mathtt{ReLU}(s_1 x) + b'_\eta$, which cannot match $\mathtt{abs}$, hence $s_2 = -s_1$. Without loss of generality (up to a permutation of indices of the hidden layer) $s_1 = 1$, $s_2 = -1$. Now, for $x < 0$ we have $-x = |x| = \boldsymbol{R}_{\theta'}(x) = -\alpha_2 x + b'_\eta$ while for $x > 0$ we get $x = |x| = \boldsymbol{R}_{\theta'}(x) = \alpha_1 x + b_\eta$, hence $\alpha_1 = 1$, $\alpha_2 = 1$, $b'_\eta = 0$. Overall, up to the possible permutation of the hidden layer, we obtain $\mathtt{sign}(\theta') = \mathtt{sign}(\theta)$ and $\alpha_1 s_1 = 1$, $\alpha_2 s_2 = -1$, $\alpha_1\beta_1 = \alpha_2\beta_2 = 0$, $b'_\eta = 0$, hence $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$. Since θ is admissible, it follows by Theorem 1 that $\theta' \sim_{PS} \theta$. Since this holds for any θ' such that $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta$, this shows that θ is PS-identifiable from $\mathcal{X} = \mathbb{R}$ with respect to $\Theta = \mathbb{R}^{E \cup \bar{H}}$.

θ is locally S-identifiable from some finite set $F \subset \mathbb{R}$ (with $0 \in F$)

With the same notations as above, observe that there is $\epsilon > 0$ such that for every $\theta' \in B(\theta, \epsilon)$ we have $s_i := \mathtt{sign}(w'_{\mu \to \nu_i}) = \mathtt{sign}(w_{\mu \to \nu_i})$, $i = 1, 2$, and $\max(|\alpha_1 - 1|, |\alpha_2 - 1|, |\beta_1|, |\beta_2|, |b'_\eta|) \leq 1/2$. Consider $\theta' \in B(\theta, \epsilon)$ such that $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_{\theta'}$ on $F = \{-3, -2, -1, 0, 1, 2, 3\}$. We have $s_1 = +1$, $s_2 = -1$ hence

$$\boldsymbol{R}_{\theta'}(x) = \alpha_1 \mathtt{ReLU}(x - \beta_1) + \alpha_2 \mathtt{ReLU}(-x - \beta_2) + b'_\eta.$$

Since $|\beta_i| \leq 1/2$ for $i = 1, 2$, we have $\boldsymbol{R}_{\theta'}(x) = \alpha_1(x - \beta_1) + b'_\eta$ for every $x \geq 1/2$, hence $\alpha_1(y - x) = \boldsymbol{R}_{\theta'}(y) - \boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_\theta(y) - \boldsymbol{R}_\theta(x) = |y| - |x| = y - x$ for $(x, y) = (1, 2)$. Therefore $\alpha_1 = 1$. A similar reasoning with $(x, y) = (-2, -1)$ shows that $\alpha_2 = 1$, hence

$$\boldsymbol{R}_{\theta'}(x) = \texttt{ReLU}(x - \beta_1) + \texttt{ReLU}(-x - \beta_2) + b'_\eta$$

Specializing to $x = 1$, since $x - \beta_1 > 0$ and $-x - \beta_2 < 0$ we get $1 = |x| = \boldsymbol{R}_{\theta'}(x) = x - \beta_1 + b'_\eta = 1 - \beta_1 + b'_\eta$ hence $b'_\eta = \beta_1$. Similarly, with $x = -1$, we get $b'_\eta = \beta_2$ hence

$$\boldsymbol{R}_{\theta'}(x) = \texttt{ReLU}(x - b'_\eta) + \texttt{ReLU}(-x - b'_\eta) + b'_\eta.$$

Specializing to $x = 0 \in F$ yields

$$0 = |x| = \boldsymbol{R}_{\theta'}(x) = b'_\eta + 2\texttt{ReLU}(-b'_\eta) = \begin{cases} b'_\eta & \text{if } b'_\eta \geq 0 \\ -b'_\eta & \text{if } b'_\eta \leq 0 \end{cases} = |b'_\eta|$$

hence $b'_\eta = 0$. Overall we have shown that for every $\theta' \in B(\theta, \epsilon)$ such that $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_\theta$ on $F = \{-2, -1, 0, 1, 2\}$ we have $\alpha_1 = \alpha_2 = 1$, $s_1 = 1, s_2 = -1$, $\beta_1 = \beta_2 = b'_\eta = 0$. These imply $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta)$ and $\texttt{sign}(\theta') = \texttt{sign}(\theta)$ hence $\theta' \sim_S \theta$. In other words, θ is locally S-identifiable from F.

Appendix J. Details on Example 5

Here we show, as claimed in Example 5 that the parameter $\theta_0 \in \mathbb{R}^{E \cup \bar{H}}$ from Example 1 is PS-identifiable from $\mathcal{X} = \mathbb{R}$ with respect to the set Θ_0 of parameters with zero output biases. Consider an arbitrary $\theta' \in \Theta_0$. If $\boldsymbol{R}_{\theta'}(x) = \boldsymbol{R}_{\theta_0}(x) = x$ on $\mathcal{X}$ then θ' is admissible (otherwise its realization would be, up to an additive constant, proportional to a single shifted version of the ReLU, which would prevent it from being equal to $\boldsymbol{R}_\theta = \texttt{id}$) hence $w'_{\nu \to \nu_i} \neq 0$, $i = 1, 2$. Writing $\alpha_i = |w'_{\mu \to \nu_i}|\ w'_{\nu_i \to \eta}$ and $\beta_i = -b'_{\nu_i}/|w'_{\mu \to \nu_i}|$, and $s_i = \texttt{sign}(w'_{\mu \to \nu_i}) \in \{-1, +1\}$ for $i = 1, 2$, we have $\alpha_i \neq 0$ and since the output bias is zero

$$\boldsymbol{R}_{\theta'}(x) = \alpha_1 \texttt{ReLU}(s_1(x - s_1\beta_1)) + \alpha_2 \texttt{ReLU}(s_2(x - s_2\beta_2)),\ \forall x \in \mathcal{X}.$$

If we had $s_1\beta_1 \neq s_2\beta_2$, $\boldsymbol{R}_{\theta'}$ would be non-differentiable at two distinct points $s_1\beta_1, s_2\beta_2$. However $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_{\theta_0} = \texttt{id}$ is differentiable on $\mathbb{R}$, hence $s_1\beta_1 = s_2\beta_2$. It follows that $s_2\beta_1 = s_1\beta_1 = \texttt{id}(s_1\beta_1) = \boldsymbol{R}_{\theta'}(s_1\beta_1) = 0$. Since $|s_i| = 1$, we get $\beta_1 = \beta_2 = 0$ and

$$\boldsymbol{R}_{\theta'}(x) = \alpha_1 \texttt{ReLU}(s_1 x) + \alpha_2 \texttt{ReLU}(s_2 x),\ \forall x \in \mathcal{X}.$$

If we had $s_1 = s_2$, the realization would be $(\alpha_1 + \alpha_2)\texttt{ReLU}(s_1 x)$, which cannot match $\texttt{id}$ on $\mathcal{X}$, hence $s_2 = -s_1$. Without loss of generality (up to a permutation of indices of the hidden layer) $s_1 = 1$, $s_2 = -1$. Now, for $x < 0$ we have $x = \texttt{id}(x) = \boldsymbol{R}_{\theta'}(x) = -\alpha_2 x$ while for $x > 0$ we get $x = \texttt{id}(x) = \boldsymbol{R}_{\theta'}(x) = \alpha_1 x$, hence $\alpha_1 = 1$, $\alpha_2 = -1$ (and $b'_\eta = 0$ because $\theta' \in \Theta_0$). Overall, up to the possible permutation of the hidden layer, we obtain $\texttt{sign}(\theta') = \texttt{sign}(\theta_0)$ and $\alpha_1 s_1 = 1$, $\alpha_2 s_2 = -1$, $\alpha_1\beta_1 = \alpha_2\beta_2 = 0$, $b'_\eta = 0$, hence $\boldsymbol{\Phi}(\theta') = \boldsymbol{\Phi}(\theta_0)$. Since θ_0 is admissible, it follows by Theorem 1 that $\theta' \sim_{PS} \theta_0$. Since this holds for any $\theta' \in \Theta_0$ such that $\boldsymbol{R}_{\theta'} = \boldsymbol{R}_{\theta_0}$, this shows that θ is PS-identifiable from $\mathcal{X} = \mathbb{R}$ with respect to Θ_0.