Most ReLU Networks Admit Identifiable Parameters

Moritz Grillo[1] and Guido Montúfar[1,2]

[1]Max Planck Institute for Mathematics in the Sciences, Leipzig
[2]Departments of Mathematics and Statistics & Data Science, UCLA

May 21, 2026

Abstract

We study the realization map of deep ReLU networks, focusing on when a function determines its parameters up to scaling and permutation. To analyze hidden redundancies beyond these standard symmetries, we introduce a framework based on weighted polyhedral complexes. Our main result shows that for every architecture whose input and hidden layers have width at least two, there exists an open set of identifiable parameters. This implies that the functional dimension of every such architecture is exactly the number of parameters minus the number of hidden neurons. We further show that minimal functional representations can still have non-trivial parameter redundancies. Finally, we establish a generic depth hierarchy, whereby for an open set of parameters the realized function cannot be represented generically by any shallower network.

Keywords: realization map, functional dimension, hidden symmetries, depth separation, minimal architecture, weighted polyhedral complex, bent hyperplane, linear regions

1 Introduction

A central question in the study of deep neural networks is the relationship between the parameter space and the space of functions they realize. For a given architecture $\mathcal{A}$, these spaces are linked by the *realization map* $\mu_{\mathcal{A}}\colon \Theta_{\mathcal{A}} \to \mathcal{F}_{\mathcal{A}};\ \theta \mapsto f_\theta$, which assigns to each parameter θ, consisting of the weights and biases of the network, the function f_θ computed by the network. The geometry of this map plays a fundamental role, as it shapes the loss landscape and its symmetries, governs implicit biases of gradient-based optimization, and determines when distinct parameter configurations represent genuinely different predictors.

A key aspect of this relationship is *identifiability*, namely whether a parameter can be uniquely recovered (up to trivial symmetries) from the function it realizes. The realization map is generally not injective due to inherent redundancies, including well-known *trivial symmetries* such as *permutations* of neurons within a layer and *positive scalings* of weights across layers. Factoring out these symmetries, we define the quotient parameter space $\overline{\Theta_{\mathcal{A}}} = \Theta_{\mathcal{A}}/\sim$. A parameter θ is said to have *no hidden symmetries*, or to be *identifiable*, if its fiber in the quotient space consists of a single element.

Parameter identifiability is closely related to the *functional dimension*, defined as the maximal rank of the Jacobian of the realization map evaluated over finite input sets:

$$\dim(\mu_{\mathcal{A}}(\Theta_{\mathcal{A}})) = \max_{\theta, X} \operatorname{rank}\left(\frac{\partial f_\theta(X)}{\partial \theta}\right),$$

where $X = \{x_1, \dots, x_N\}$ is a finite input set and $f_\theta(X)$ denotes the evaluation of the realized function on X. Intuitively, the functional dimension is the maximal number of independent directions in function space that can be realized by infinitesimal variations of the parameters. If the realization map is smooth on a nonempty open set of identifiable parameters, then it is locally injective modulo the trivial symmetries, and the architecture attains its *expected functional dimension*, given by $\dim(\mu_{\mathcal{A}}(\Theta_{\mathcal{A}})) = \dim(\overline{\Theta_{\mathcal{A}}})$.

Previous work has established that certain classes of architectures attain the expected functional dimension. In particular, Phuong and Lampert (2020) proved for pyramidal networks with decreasing layer widths that there is an open set of parameters that are identifiable among generic parameters, while Grigsby et al. (2023), building on Rolnick and Kording (2020), established the expected functional dimension for architectures whose hidden layers have width at least the input dimension $n_i \geq d \geq 2$. However, a general structural understanding of identifiability for ReLU architectures remains lacking, raising the following question:

Which ReLU architectures attain the expected functional dimension?

Factoring out trivial symmetries, one may ask whether any remaining non-identifiability can be explained by non-minimality of the architecture. In many settings, non-identifiability can be linked to a lack of *minimality*, meaning that the function can be realized by a strictly smaller sub-architecture, or *submodel* (Sussmann, 1992, Fukumizu and Amari, 2000a, Petzka et al., 2020, Shahverdi et al., 2026a). Such submodels induce redundancy through overparameterization, as parameters corresponding to redundant neurons can be varied without affecting the realized function. This raises the following fundamental question for deep ReLU networks:

Does non-identifiability in ReLU networks arise solely from reducibility of the architecture, or can additional forms of redundancy persist even for minimal realizations?

The difficulty of these questions is compounded by the structure of ReLU networks. The ReLU activation is not differentiable at zero and vanishes on a half-space, leading to realization maps whose Jacobians, evaluated over a finite input set, depend intricately on the network's activation pattern. These patterns are determined by semi-algebraic regions in parameter space whose structure remains poorly understood. As a result, characterizing the rank of the Jacobian is highly nontrivial. This problem has been studied extensively in the context of the empirical neural tangent kernel (NTK), defined as the Gram matrix of these Jacobians. However, such works focus on showing that sufficiently overparametrized networks attain full-rank NTK with high probability over a random parameter. In contrast, the functional dimension is an intrinsic property of the architecture, requiring maximization of the Jacobian rank jointly over the parameters and input datasets.

In this work, we address these questions by developing a unified framework for general ReLU architectures. We show that all deep ReLU architectures of width at least two attain the expected functional dimension. Moreover, such architectures can simultaneously admit open sets of identifiable parameters and open sets of minimal parametrizations that are not identifiable, revealing intrinsic redundancies beyond standard symmetries and submodel structure.

1.1 Our Contributions

In this paper, we address fundamental questions regarding parameter identifiability of deep ReLU networks. Our contributions can be summarized as follows:

1. **A Unified Polyhedral Framework:** In Section 3, we introduce a framework to study identifiability and parameter fibers, describing functional non-linearities via weighted polyhedral complexes.

2. **Identifiability in Almost all Architectures:** In Section 4, we show that for any architecture whose input and hidden layers have width at least two, there exists a nonempty open subset of parameters that are identifiable among *all* parameters. This substantially generalizes prior genericity and width-dependent results of Grigsby et al. (2023) and Phuong and Lampert (2020).

3. **Functional Dimension:** As a consequence, we settle the functional dimension for nearly all ReLU architectures as the number of parameters minus the number of hidden neurons (Theorem 4.11).

4. **Minimality $\neq$ Identifiability:** In Section 5, we construct an open set of minimal parameters with positive-dimensional non-trivial fibers. This demonstrates that redundancy can persist even when no neurons can be removed.

5. **Generic Depth Hierarchy:** In Section 6, we show that, for most architectures, there exists an open set of parameters whose realized functions cannot be realized by any shallower network with generic parameters, regardless of width.

1.2 Related Work

Symmetries and Identifiability The study of functional equivalence and minimal realizations in neural networks has a long history. Early work by Sussmann (1992) studies identifiability and minimality for sigmoidal networks, showing that non-identifiability can arise from non-minimal realizations and providing conditions under which minimal representations are unique. In a related direction, Kůrková and Kainen (1994) investigate when two feedforward networks represent the same function, characterizing equivalence classes of parameters and highlighting the role of redundancy in network representations. Fefferman (1994) gives conditions under which knowledge of the output map determines the architecture and parameters up to standard symmetries in the case of tanh activation. Moreover, Vlačić and Bölcskei (2021) develop a general framework in which fibers are generated by affine symmetries of the activation function, giving an exhaustive description except in the presence of nontrivial zero-realizing networks.

A substantial body of work has examined the impact of parameter symmetries on optimization, from the geometry of loss landscapes in hierarchical structures (Fukumizu and Amari, 2000b, Simsek et al., 2021) to the topology of training dynamics (Nurisso et al., 2026). A recent survey on parameter symmetries is provided by Zhao et al. (2025).

We next highlight a line of work on neural network identifiability in settings where the realization map admits an algebraic structure. For polynomial models, classical results on the dimension of secant varieties (Alexander and Hirschowitz, 1995) yield functional dimension results for shallow polynomial networks. For deep networks with polynomial activations, recent works (Shahverdi et al., 2026b, Usevich et al., 2025, Finkel et al., 2025) study generic identifiability and expected dimension using tools from algebraic geometry, showing that sufficiently high activation degree ensures identifiability up to permutation and scaling symmetries. Recent works have started to classify the parameter symmetries in transformers (Tran et al., 2025).

We also highlight a line of work on dimension and finite identifiability for restricted Boltzmann machines (RBMs). These are energy-based probabilistic models whose log-probabilities correspond to shallow softplus networks, while their tropicalizations correspond to shallow ReLU networks, both with unit output weights and evaluated on binary inputs (Montúfar, 2018). Cueto et al. (2010) showed that many tropicalized RBM architectures attain the expected dimension, and later Montúfar and Morton (2017) established that the original model always does. However, the dimension of the tropical RBM, and thus of the corresponding shallow ReLU network on binary inputs, remains only partially understood. We next review works specific to ReLU networks.

Identifiability and Dimension of ReLU Networks For ReLU networks, identifiability is inherently combinatorial and geometric due to the piecewise linear activation. For shallow ReLU networks, Petzka et al. (2020) and Ramakrishnan (2026) provide a complete characterization of symmetries in the two-layer case, identifying redundancies beyond permutation and scaling. Moreover, Dereich and Kassing (2022) characterize minimal representations of functions realized by shallow ReLU networks and determine both the dimension and the number of connected components of the corresponding function space. Building on this perspective, Dereich and Kassing (2024) show that optimal approximations of continuous target functions exist within this space. Our work builds on the reverse-engineering framework of Rolnick and Kording (2020), which exploits the geometry of bent hyperplanes and activation boundaries to reconstruct network parameters from the realized function.

Following this line, Grigsby et al. (2023) establish the existence of open sets of parameters satisfying geometric conditions under which the reverse-engineering framework applies for architectures whose hidden layers have width at least the input dimension, thereby showing that these architectures attain the expected functional dimension. Likewise, Phuong and Lampert (2020) prove the existence of an open set of parameters that are identifiable among generic parameters for pyramidal architectures with non-increasing widths, provided that all input and hidden layers have width at least two. For pyramidal architectures, Bona-Pellissier et al. (2023) further provide conditions under which equality of realized functions implies equality of parameters up to trivial symmetries, and exhibit examples that fall outside the cases covered by Phuong and Lampert (2020). Moreover, Stock and Gribonval (2023) introduce a locally linear parametrization of the realization map in terms of paths, derive conditions for local identifiability, and provide necessary and sufficient conditions in the shallow case.

A related notion is that of functional dimension. For ReLU networks, Grigsby et al. (2025) show that it varies across parameter space, while Grigsby and Lindsey (2024) relates this local dimension to the persistent pseudodimension as a measure of local capacity. Complementarily, the constraints of the realizable functions across regions of parameter space have recently been investigated in Alexandr and Montúfar (2025).

Finally, we note that, closely related to the functional dimension, an extensive line of works has studied the degree of overparametrization required for the NTK to attain full rank with high probability. Several works have established lower bounds on the smallest eigenvalue (Nguyen et al., 2021, Bombari et al., 2022, Montanari and Zhong, 2022, Karhadkar et al., 2024a). Focusing on rank, Karhadkar et al. (2024b) show that for mildly overparametrized networks most activation patterns correspond to parameter regions where the Jacobian attains full rank equal to the number of data points.

Linear Regions and Depth Separation There is substantial work that studies subdivisions of the input space induced by linear regions and their connection to depth separation. One direction studies approximation-theoretic separation between shallow and deep networks, including the results of Eldan and Shamir (2016), Telgarsky (2016), and Mhaskar et al. (2017), as well as works quantifying complexity through the proliferation of linear regions (Montúfar et al., 2014, Pascanu et al., 2014, Raghu et al., 2017, Serra et al., 2018, Balestriero et al., 2019, Ergen and Grillo, 2024). These results demonstrate that depth enhances expressive power, as reflected in approximation rates, oscillatory behavior, and exponential growth in the number of linear regions. More recent works in this direction provide explicit counting formulas by relating linear regions to Minkowski sums of polytopes (Montúfar et al., 2022), and recent works aim to obtain precise structural characterization of the associated polytopes (Balakin et al., 2025).

A complementary line of work focuses on depth hierarchies for exact representation of continuous piecewise linear functions by ReLU networks irrespective of width. This perspective was initiated by Hertrich et al. (2023) via the representability of the maximum function and related maps. Subsequent works establish lower bounds in several regimes: Haase et al. (2023) prove logarithmic depth lower bounds for networks with integer weights, Averkov et al. (2025) extend this to rational-weights, and Grillo et al. (2025) derive lower bounds under compatibility assumptions with the braid arrangement. At the same time, Bakaev et al. (2026) show that earlier conjectured upper bounds for exact depth hierarchies are not tight by constructing shallower representations of the maximum function. Our generic depth-hierarchy result is complementary to this line: rather than focusing on specific functions such as max, it shows that for generic parameters depth cannot be traded for width.

Weighted Polyhedral Complexes for Neural Networks Polyhedral and tropical perspectives have become central tools for understanding ReLU networks and, more generally, continuous piecewise-linear models. Zhang et al. (2018) and Charisopoulos and Maragos (2018) identify ReLU networks with tropical rational maps, relating their linear regions and decision boundaries to polyhedral and tropical-geometric structures. This viewpoint has since been refined in several directions. For instance, Brandenburg et al. (2024) study semialgebraic subdivisions of parameter space and the combinatorial types of decision boundaries, highlighting that parameter space itself admits a rich stratification. A broader overview of polyhedral methods in deep learning was recently presented by Huchette et al. (2023).

Closer to our setting, weighted polyhedral complexes have been used to encode function representations and decompositions. In particular, Tran and Wang (2024) study minimal representations of tropical rational functions, directly addressing questions of redundancy and factorization length for piecewise-linear maps. In a similar spirit, Brandenburg et al. (2025) introduce decomposition polyhedra for CPWL functions, showing that, once an underlying complex is fixed, the space of admissible decompositions is organized as a polyhedron whose minimal decompositions are vertices. This viewpoint is closely aligned with our use of weighted polyhedral complexes to separate visible function geometry from latent parameter geometry.

2 Preliminaries

Notation For any $n \in \mathbb{N}$, we let $[n] := \{1, \dots, n\}$. For a vector $x \in \mathbb{R}^d$, we denote by $[x]_+$ the entrywise application of the ReLU activation function $\max\{0, x_i\}$. For a set of indices $S \subseteq [n]$, let $D_S \in \mathbb{R}^{n \times n}$ denote the diagonal selection matrix where $(D_S)_{ii} = 1$ if $i \in S$ and $(D_S)_{ii} = 0$ otherwise. We also simply write $\|\cdot\|$ for the euclidean norm $\|\cdot\|_2$.

ReLU Networks A *ReLU layer* with $n_{\ell-1}$ inputs and n_ℓ outputs, weight matrix $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, and bias vector $b^{(\ell)} \in \mathbb{R}^{n_\ell}$ computes the map $f_{W^{(\ell)}, b^{(\ell)}}(x) = [W^{(\ell)}x + b^{(\ell)}]_+$.

A *ReLU network* with architecture $\mathcal{A} = (n_0, \dots, n_{L+1})$ is the composition

$$f_\theta(x) = T_{W^{(L+1)}, b^{(L+1)}} \circ f_{W^{(L)}, b^{(L)}} \circ \cdots \circ f_{W^{(1)}, b^{(1)}}(x),$$

where the final layer $T_{W^{(L+1)}, b^{(L+1)}}(y) = W^{(L+1)}y + b^{(L+1)}$ is an affine linear map. We sometimes write d for the input dimension n_0.

The tuple $\theta = (W^{(1)}, b^{(1)}, \dots, W^{(L+1)}, b^{(L+1)})$ is called the *parameter* of the network. The *parameter space* is given by $\Theta_{\mathcal{A}} \cong \bigoplus_{\ell=1}^{L+1} \Theta_\ell$, where $\Theta_\ell = \mathbb{R}^{n_\ell \times n_{\ell-1}} \times \mathbb{R}^{n_\ell}$. The *realization map* $\mu_{\mathcal{A}} \colon \Theta_{\mathcal{A}} \to \mathcal{F}_{\mathcal{A}}$ is defined by $\theta \mapsto f_\theta$. For $\ell \in [L]$, the *preactivation* at layer ℓ is a function of the network's input defined by $z^{(\ell,\theta)}(x) = W^{(\ell)}a^{(\ell-1,\theta)}(x) + b^{(\ell)}$, where $a^{(\ell,\theta)}(x) := [z^{(\ell,\theta)}(x)]_+$ is the *activation* at layer ℓ and $a^{(0,\theta)} = x$. The tuple (j, ℓ) indexes a *neuron in layer* ℓ with preactivation $z_j^{(\ell)}$ and activation $a_j^{(\ell)}$. The j-th neuron in layer ℓ has input weights given by $W_j^{(\ell)}$, the j-th row, and output weights given by $W_{:,j}^{(\ell+1)}$, the jth column of $W^{(\ell+1)}$.

We also use the notation $\Theta^{(\ell)} = \bigoplus_{k=1}^{\ell} \Theta_k$ for the parameter space of the first ℓ layers. For a parameter $\theta = (\theta_1, \dots, \theta_{L+1}) \in \Theta_{\mathcal{A}}$, we use the notation $\theta^{(\ell)} = (\theta_1, \dots, \theta_\ell) \in \Theta^{(\ell)}$ and denote by $f_{\theta^{(\ell)}}$ the corresponding truncated network with ReLU applied to the output layer for $\ell \le L$. Note that the functions $f_{\theta^{(\ell)}}$ and $a^{(\ell,\theta)}$ are identical, but the distinguished notation will become handy in Section 4.

Symmetries, Identifiability and Minimality The parameter space $\Theta_{\mathcal{A}}$ carries a natural equivalence relation, denoted $\theta \sim \eta$, generated by the *trivial symmetries* of the networks (see, e.g., Rolnick and Kording, 2020, Phuong and Lampert, 2020):

1. The neurons within any layer $\ell \in [L]$ can be reordered without altering the realized function f_θ.
2. For any neuron j in layer ℓ and any scaling factor $\lambda > 0$, one can multiply the j-th row of $W^{(\ell)}$ and the bias entry $b_j^{(\ell)}$ by λ, provided the j-th column of $W^{(\ell+1)}$ is multiplied by λ^{-1}.

We denote by $\overline{\Theta_{\mathcal{A}}} = \Theta_{\mathcal{A}}/\sim$ the parameter space modulo these trivial symmetries. A parameter θ is *identifiable* if its fiber consists only of its symmetry orbit, i.e., $f_\eta = f_\theta$ implies $\eta \sim \theta$. In terms of the realization map, θ is identifiable if its image in the quotient space has a trivial fiber: $\overline{\mu}_{\mathcal{A}}^{-1}(f_\theta) = \{[\theta]\}$. By abuse of notation, we will often identify a parameter $\theta \in \Theta_{\mathcal{A}}$ with its equivalence class $[\theta] \in \overline{\Theta_{\mathcal{A}}}$ whenever we work in the quotient space. Moreover, if $\mathcal{A}' = (n_0, n_1', \dots, n_L', n_{L+1})$ satisfies $n_\ell' \le n_\ell$ for all ℓ, then $\mathcal{A}'$ is called a *subarchitecture* or *submodel* of $\mathcal{A}$, and θ is *minimal* if no strict submodel can realize the same function.

Polyhedral Geometry For a vector $a \in \mathbb{R}^d$ and a scalar $b \in \mathbb{R}$, the *hyperplane* $H := \{x \in \mathbb{R}^d \mid \langle a, x \rangle + b = 0\}$ subdivides $\mathbb{R}^d$ into *half-spaces* $H^+ := \{x \in \mathbb{R}^d \mid \langle a, x \rangle + b \ge 0\}$ and $H^- := \{x \in \mathbb{R}^d \mid \langle a, x \rangle + b \le 0\}$. A finite set of hyperplanes $\mathcal{H} = \{H_1, \dots, H_n\}$ is called a *hyperplane arrangement*. A hyperplane arrangement $\mathcal{H}$ in $\mathbb{R}^d$ is called *generic* if the intersection of any subset of $k \le d$ hyperplanes has codimension k, and no

$d+1$ hyperplanes have a common intersection. Given an arbitrary arrangement $\mathcal{H}$, let $L \coloneqq \bigcap_{H \in \mathcal{H}} \operatorname{lin}(H)$ be the maximal linear subspace contained in all hyperplanes. Projecting $\mathbb{R}^d$ orthogonally onto $L^\perp$ yields an induced arrangement $\mathcal{H}^{\mathrm{ess}}$ in $L^\perp$, called the *essentialization* of $\mathcal{H}$.

A *polyhedral complex* $\mathcal{C}$ is a finite collection of polyhedra such that

1. $\emptyset \in \mathcal{C}$,
2. if $P \in \mathcal{C}$, then all faces of P are in $\mathcal{C}$, and
3. if $P, P' \in \mathcal{C}$ and $P \cap P' \neq \emptyset$, then $P \cap P'$ is a face of both P and P'.

For a polyhedral complex $\mathcal{C}$ in $\mathbb{R}^d$ and $k \leq d$ we denote by $\mathcal{C}^k$ the set of k-dimensional polyhedra in $\mathcal{C}$. We call $\mathcal{C}^{d-1}$ the *facets*, $\mathcal{C}^d$ the *regions* and $\mathcal{C}^{d-2}$ the *ridges* of $\mathcal{C}$. The *dimension* of a complex $\mathcal{C}$ is the maximal dimension of its polyhedra. A complex is *pure* (of dimension k) if all maximal polyhedra are of dimension k. A pure polyhedral complex in $\mathbb{R}^d$ of dimension $d-1$ is called a *piecewise linear hypersurface*. A *local normal* at $x \in \mathbb{R}^d$ of a piecewise linear hypersurface is a vector that is orthogonal to one of the maximal polyhedra containing x.

Given a face $\sigma \in \mathcal{C}$, we denote by $\operatorname{aff}(\sigma) \subseteq \mathbb{R}^d$ the unique smallest affine subspace containing σ. The *relative interior* of σ is the interior of σ inside the affine space $\operatorname{aff}(\sigma)$ and is denoted by $\operatorname{relint}(\sigma)$.

Let $\mathcal{C}$ be a polyhedral complex in $\mathbb{R}^d$ and let $\tau \in \mathcal{C}$ be a face. The *star* of τ is $\operatorname{star}_{\mathcal{C}}(\tau) \coloneqq \{ \sigma \in \mathcal{C} \mid \tau \subseteq \sigma \}$. When $\mathcal{C}$ is clear from the context, we omit the subscript and write $\operatorname{star}(\tau)$. For any $k \leq d$ and any faces $\tau \in \mathcal{C}^{k-1}, \sigma \in \mathcal{C}^k$ with $\tau \subseteq \sigma$, let $e_{\sigma/\tau} \in \mathbb{R}^d$ denote the normal vector of τ relative to σ, defined as the unique unit vector that lies in $\operatorname{aff}(\sigma)$, is orthogonal to $\operatorname{aff}(\tau)$, and points from the relative interior of τ into the relative interior of σ. For a subset $S \subseteq \mathcal{C}$, we denote the *support* by $|S| \coloneqq \bigcup_{P \in S} P$ and by $\#S$ the number of elements contained in S.

A function $f \colon \mathbb{R}^d \to \mathbb{R}^m$ is *continuous and piecewise linear* (CPWL), if there exists a complete polyhedral complex $\mathcal{C}$ such that the restriction of f to each polyhedron $P \in \mathcal{C}$ is an affine linear function. If this condition is satisfied, we say that f and $\mathcal{C}$ are *compatible* with each other. A vector $x \in \mathbb{R}^d$ is a *breakpoint* of f if there is no open set $U \subseteq \mathbb{R}^d$ containing x such that f is affine linear on U. For a CPWL function $f \colon \mathbb{R}^d \to \mathbb{R}^m$, let $B(f)$ be the set of breakpoints of f.

A polyhedral complex $\mathcal{C}$ in $\mathbb{R}^d$ of dimension at least $d-1$ can be equipped with a *weight function* $c \colon \mathcal{C}^{d-1} \to \mathbb{R}^m$. One can use such a weight function to describe a CPWL function by storing how the linear parts change when crossing a facet in a compatible complex as formalized in the following lemma. See also Figure 1a for an illustration.

Lemma 2.1. *Let $f \colon \mathbb{R}^d \to \mathbb{R}^m$ be a CPWL function compatible with a polyhedral complex $\mathcal{C}$. For a facet $\sigma \in \mathcal{C}^{d-1}$, let $P, Q \in \mathcal{C}^d$ be the unique polyhedra such that $P \cap Q = \sigma$, and suppose that $f(x) = A_P x + b_P$ for all $x \in P$ and $f(x) = A_Q x + b_Q$ for all $x \in Q$. Then the tropical weight $c_f \colon \mathcal{C}^{d-1} \to \mathbb{R}^m$ defined by*

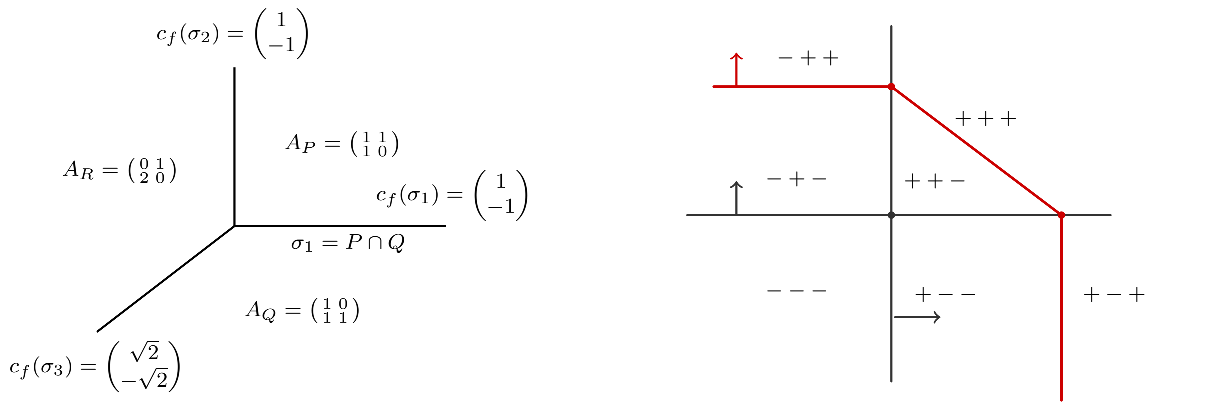
$$c_f(\sigma) \coloneqq A_P e_{P/\sigma} + A_Q e_{Q/\sigma} = (A_P - A_Q) e_{P/\sigma}$$

uniquely determines f up to adding a global linear function.

Proof. Write $f = (f_1, \ldots, f_m)$ with $f_i \colon \mathbb{R}^d \to \mathbb{R}$. Since f is CPWL and compatible with the polyhedral complex $\mathcal{C}$, each coordinate function f_i is CPWL and compatible with the same complex. Then the result follows from Tropical Geometry (Maclagan and Sturmfels, 2015, Proposition 3.3.10 and Maclagan and Sturmfels, 2015, Proposition 3.3.2), as outlined by Brandenburg et al. (2025). ☐

3 A Polyhedral Framework for Identifiability

In this section we develop the polyhedral language that will be used throughout the paper to study identifiability. The main idea is to separate two types of structure associated with a ReLU network: the *canonical polyhedral complex*, which records all sign changes of neuron preactivations layer by layer, and the *breakpoint*



(a) Illustration of a weighted polyhedral complex. The matrices on the maximal polyhedra represent the linear parts of f, and the labels on the facets show how the tropical weights record the magnitude of the change of the linear part across the facets.

(b) Illustration of the canonical polyhedral complex. The black hyperplanes represent the zero-loci of two neurons in the first hidden layer and the red bent hyperplane the zero-locus of a neuron in the second hidden layer. The sign sequences index polyhedra of the complex.

Figure 1: Illustration of weighted polyhedral complexes, the canonical polyhedral complex and bent hyperplanes.

complex, which records only those pieces of this geometry that remain visible in the realized function. We first introduce the canonical complex and bent hyperplanes, then identify a generic regime in which their geometry is well behaved, and finally show how the local weighted geometry of the breakpoint complex gives rise to the dependency graph that encodes layer precedence. Finally, we establish sufficient conditions for a parameter to be *identifiable among generic parameters*. Some of the concepts introduced along the way have appeared, in various forms, in prior work (Grigsby et al., 2025, Masden, 2025, Phuong and Lampert, 2020, Rolnick and Kording, 2020). The main goal of this section is to develop a unified polyhedral language for these concepts to rigorously state a sufficient condition for identifiability.

3.1 Canonical Polyhedral Complex and Bent Hyperplanes

In this subsection, following Grigsby and Lindsey (2022), we introduce the canonical polyhedral complex and the associated bent hyperplanes. The canonical complex records the activation regions of the hidden neurons, while bent hyperplanes isolate the loci where individual neurons change their affine behavior.

Canonical Polyhedral Complex Let $\phi_{W,b}\colon \mathbb{R}^d \to \mathbb{R}^m$ be a ReLU layer with weight matrix $W \in \mathbb{R}^{m\times d}$ and bias vector $b \in \mathbb{R}^m$. The breakpoints of $\phi_{W,b}$ are contained in the hyperplane arrangement $\mathcal{H}_{W,b} \coloneqq \{\{x \in \mathbb{R}^d \mid W_i x + b_i = 0\}\}_{i=1}^m$. Each sign pattern $\mathbf{s} = (s_1, \ldots, s_m) \in \{+, 0, -\}^m$ determines a (possibly empty) polyhedron $P_{\mathbf{s}} \coloneqq \bigcap_{i=1}^m H^{s_i}_{W_i,b_i}$, where H^+ and H^- denote the corresponding half-spaces and H^0 the hyperplane. The collection of all such polyhedra forms the *canonical polyhedral complex* $\mathcal{C}_{W,b}$ of $\phi_{W,b}$. On each $P_{\mathbf{s}}$, the map $\phi_{W,b}$ is affine linear. If $\mathcal{H}$ is a hyperplane arrangement given by W and b, then we also write $\mathcal{C}_{\mathcal{H}}$ for $\mathcal{C}_{W,b}$.

Let $\theta \in \Theta_{\mathcal{A}}$ for $\mathcal{A} = (n_0, \ldots, n_{L+1})$. The canonical polyhedral complex associated with θ is constructed iteratively, layer by layer. We start with the trivial complex $\mathcal{C}_{\theta,0} \coloneqq \mathbb{R}^d$. Suppose that at stage $\ell - 1$ we have a polyhedral complex $\mathcal{C}_{\theta,\ell-1}$ such that all preactivations up to layer ℓ are affine linear on each polyhedron $R \in \mathcal{C}_{\theta,\ell-1}$. For each such polyhedron R and each neuron (ℓ, j), the preactivation $z_j^{(\ell)}$ restricts to an affine map on R. If $z_j^{(\ell)}|_R(x)$ is not constant, then its zero set defines a hyperplane $H_R(\ell, j) \coloneqq \{x \in \operatorname{aff}(R) \mid z_j^{(\ell)}|_R(x) = 0\}$ which coincides with the breakpoint set introduced by neuron (ℓ, j) on R. If $z_j^{(\ell)}|_R(x)$ is constant, then

$H_R(\ell, j)$ is simply $\operatorname{aff}(R)$ or the empty set, in which case it does not affect the further refinement described below. We refine the complex $\mathcal{C}_{\theta,\ell-1}$ by subdividing each R using the hyperplane arrangement induced by layer ℓ on $\operatorname{aff}(R)$:

$$\mathcal{C}_{\theta,\ell} := \{\, R \cap \bigcap_{j=1}^{n_\ell} H_R^{s_j}(\ell, j) \mid R \in \mathcal{C}_{\theta,\ell-1},\ \mathbf{s} \in \{+, 0, -\}^{n_\ell} \,\}.$$

We define the *canonical polyhedral complex* of θ as $\mathcal{C}_\theta := \mathcal{C}_{\theta,L}$. Consequently, the polyhedra of $\mathcal{C}_\theta$ are indexed by global activation patterns, that is, $\mathcal{C}_\theta = \{P_\mathbf{s} \mid \mathbf{s} \in \{+, 0, -\}^{n_1} \times \cdots \times \{+, 0, -\}^{n_L}\}$, where each polyhedron $P_\mathbf{s} := \overline{\{x \in \mathbb{R}^d \mid \operatorname{sign} z_j^{(\ell)}(x) = \mathbf{s}_{\ell,j}, \ell \in [L], j \in [n_\ell]\}}$ is the closure of all inputs realizing the corresponding signs of all neuron preactivations. For $P \in \mathcal{C}_\theta$, we also write $\mathbf{s}_{\ell,j}(P) = \operatorname{sign} z_j^{(\ell)}(x)$ for any $x \in \operatorname{relint}(P)$. On every such polyhedron, the function f_θ restricts to an affine-linear map.

For $S = (S_1, \cdots, S_\ell)$ with $S_k \subseteq [n_k]$, let

$$P(S) := \overline{\{x \in \mathbb{R}^d \mid \operatorname{sign} z_j^{(k,\theta)}(x) = + \text{ iff } j \in S_k \text{ for all } k \in [\ell], j \in [n_k]\}}.$$

All maximal polyhedra in $\mathcal{C}_{\theta,\ell}$ are of this form. For $P \in \mathcal{C}_\theta$, we denote by $\mathbf{s}(P)$ the corresponding sign pattern and by $S(P) = (S_1(P), \ldots, S_L(P))$ the set of active neurons, that is, $j \in S_\ell(P)$ if and only if $\mathbf{s}_{\ell,j}(P) = +$ for all $j \in [n_\ell]$.

We also need to describe the geometry locally, motivating the following definition. Let $P \subseteq \mathbb{R}^d$ be a polyhedron and $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^m$. Then we define the *local canonical polyhedral complex* as $\mathcal{C}_{\theta,\ell}(P) := \{R \cap P \mid R \in \mathcal{C}_{\theta,\ell}\}$. The polyhedra in $\mathcal{C}_\theta(P)$ inherit the activation patterns from $\mathcal{C}_\theta$.

Weighted Canonical Complex A useful way to encode a CPWL function compatible with a polyhedral complex is via its *weight function*. By Lemma 2.1, if $f \colon \mathbb{R}^d \to \mathbb{R}^m$ is compatible with a polyhedral complex $\mathcal{C}$, then f determines a map $c_f \colon \mathcal{C}^{d-1} \to \mathbb{R}^m$ that records the change of the affine-linear part of f across each facet. This weight function determines f up to adding a global affine-linear map. For a neural network f_θ, we write the induced weight function on the canonical polyhedral complex $\mathcal{C}_\theta$ simply as $c_\theta \colon \mathcal{C}_\theta^{d-1} \to \mathbb{R}^m$.

Bent Hyperplanes For a neuron (j, ℓ) in layer $\ell \in [L]$, the breakpoints form a piecewise linear hypersurface. We define the *bent hyperplane* associated with the preactivation $z_j^{(\ell,\theta)}$ as:

$$B_{j,\ell}(\theta) := \bigcup_{\substack{R \in \mathcal{C}_{\theta,\ell-1} \\ z_j^{(\ell)}|_R \text{ non-constant}}} \{x \in R \mid z_j^{(\ell)}|_R(x) = 0\}.$$

The bent hyperplane $B_{j,\ell}(\theta)$ is the locus of points where the j-th neuron of layer ℓ is nonlinear. See also Figure 1b for an illustration of the canonical polyhedral complex and bent hyperplanes. This formulation of the bent hyperplane differs slightly from the one provided by Grigsby and Lindsey (2022) and Rolnick and Kording (2020), which define it as the full zero-locus of the neuron's preactivation. While that definition would technically include entire regions where the preactivation vanishes identically, our definition focuses on the locus where the neuron actually changes its linear behavior. Importantly, for generic parameters, preactivations are non-zero almost everywhere, and the two definitions coincide

Linear Regions A *linear region* of a CPWL function $f \colon \mathbb{R}^d \to \mathbb{R}^m$ is a maximal subset of $\mathbb{R}^d$ with connected interior on which f is affine-linear. In general, a linear region of f_θ may be a union of several maximal polyhedra of $\mathcal{C}_\theta$ if the affine-linear functions on adjacent polyhedra happen to coincide. We say that f_θ satisfies the *Linear Region Assumption (LRA)* on a set $X \subseteq \mathbb{R}^d$ if the linear regions of the restricted function $f_\theta|_X$ coincide exactly with the sets $P \cap X$, where P ranges over the maximal polyhedra of $\mathcal{C}_\theta$ intersecting X.

3.2 The Geometry of Generic Parameters

We now introduce geometric conditions ensuring that the canonical polyhedral complex and the bent hyperplanes are well behaved. These conditions hold for almost all parameter choices and define the generic regime used throughout the paper, which we define below.

Definition 3.1. We call a parameter $\theta \in \Theta_{\mathcal{A}}$ *generic* if it satisfies the following two conditions:

1. For every layer $\ell \in [L]$ and every region $R \in \mathcal{C}_{\theta,\ell-1}$, the essentialization of the hyperplane arrangement $\{H_R(\ell, j)\}_{j\in[n_\ell]}$ is generic and does not intersect any vertex of R.
2. For every $k, \ell \in [L+1]$ and every sequence of index sets $S_i \subseteq [n_i]$ for $k \leq i \leq \ell-1$, the matrix product

$$W^{(\ell)} D_{S_{\ell-1}} W^{(\ell-1)} \cdots D_{S_k} W^{(k)}$$

achieves the maximum possible rank, namely $\min(n_{k-1}, n_\ell, \min_{k\leq i\leq \ell-1} |S_i|)$.

We denote by $\tilde{\Theta}_{\mathcal{A}} \subseteq \Theta_{\mathcal{A}}$ the set of generic parameters.

Lemma 3.2. *Generic parameters form an open and dense subset of* $\Theta_{\mathcal{A}}$.

Proof. The set of generic parameters is the complement of a finite union of proper algebraic varieties. To see this for the first condition, fix a sequence of activation patterns. Under such an activation pattern, the zero set of each neuron defines a hyperplane whose normal and translation are polynomial functions of the network's weights and biases. Requiring that the essentialization of $\{H_R(\ell, j)\}_{j\in[n_\ell]}$ is generic and does not intersect any vertex of R, amounts to avoiding the zero set of a finite collection of polynomial equations given by the minors of matrices consisting of facet-defining normals of the region R and normals of the hyperplanes.

The second condition requires that products of weight matrices and diagonal selection matrices have rank at least $\min(n_{k-1}, n_\ell, \min_{k\leq i\leq \ell-1} |S_i|)$. The rank $\geq r$ condition fails if and only if all $r \times r$ minors vanish. Since these minors are polynomial functions of the network weights, the failure of the rank condition is characterized by the zero locus of a finite collection of polynomials.

In both cases, the exceptional sets are proper algebraic varieties. Their union is therefore closed with empty interior, and its complement is open and dense. $\square$

Definition 3.3. A parameter θ is *supertransversal* if every face $\tau \in \mathcal{C}_\theta^{d-k}$ is contained in exactly k bent hyperplanes, and *cancellation-free* if, for every facet $\sigma \in \mathcal{C}_\theta^{d-1}$, it holds that

$$c_\theta(\sigma) = 0 \quad \Longleftrightarrow \quad \text{there exists a layer } k \in [L] \text{ such that } \mathbf{s}_{k,j}(\sigma) = - \text{ for all } j \in [n_k].$$

The concept of supertransversality was introduced in an early version in the work of Masden (2025) in a stronger form, whose defining conditions imply those in the definition given above. See Figure 2c for an illustration of a situation where supertransversality fails. Cancellation-freeness, without using this specific name, is a property that has already been used by Phuong and Lampert (2020).

We will need the following standard lemma which we include for completeness.

Lemma 3.4. *Let $\mathcal{H}$ be a hyperplane arrangement whose essentialization is generic, and let τ be a codimension-k face of the polyhedral complex induced by $\mathcal{H}$. Then τ is contained in exactly k hyperplanes of $\mathcal{H}$.*

Proof. Let L be the lineality space of $\mathcal{H}$. Passing to the essentialization $\mathcal{H}^{\text{ess}}$ in $L^\perp$ preserves face inclusions and intersections. Since $\mathcal{H}^{\text{ess}}$ is a generic arrangement, any codimension-k face in $\mathcal{H}^{\text{ess}}$ is formed by the intersection of exactly k hyperplanes. Lifting this back to $\mathbb{R}^d$, τ must also be contained in exactly k hyperplanes of $\mathcal{H}$. $\square$

Lemma 3.5. *Generic parameters are supertransversal.*

Proof. We prove the statement by induction on the layers of the network. For the first layer, the bent hyperplanes are simply the affine hyperplanes $H_1, \dots, H_{n_1}$. Since θ is generic (Definition 3.1), the essentialization of this hyperplane arrangement is generic (Condition 1 of Definition 3.1). Hence, by Lemma 3.4, any codimension-k face $\tau \in \mathcal{C}_{\theta,1}^{d-k}$ lies in exactly k hyperplanes. This establishes the base case.

Assume that for the first $\ell-1$ layers, every codimension-k face $\tau \in \mathcal{C}_{\theta,\ell-1}^{d-k}$ lies in exactly k bent hyperplanes from $\ell-1$ layers. Let $\tau \in \mathcal{C}_{\theta,\ell}^{d-k}$ be a codimension-k face in the refined complex obtained after adding layer ℓ. By construction, τ arises as the intersection of some polyhedron $R \in \mathcal{C}_{\theta,\ell-1}$ with a collection of half-spaces or hyperplanes

$$\{H_R^{s_j}(\ell, j)\}_{j\in[n_\ell],\, s_j\in\{+,-,0\}}$$

induced by the neurons in layer ℓ. Let k' denote the codimension of R in $\mathbb{R}^d$. By the induction hypothesis, R is contained in exactly k' bent hyperplanes from the first $\ell-1$ layers.

If $k' = d$ (so that $\dim(R) = 0$), then $\tau = R$ is a vertex. Since θ is generic, no hyperplane from layer ℓ passes through a vertex of R (Condition 1 of Definition 3.1). Hence, no additional intersections with layer-ℓ hyperplanes occur, and τ is contained in exactly $k' = k$ bent hyperplanes.

Otherwise, R has codimension $k' < d$ in $\mathbb{R}^d$, and the codimension of $\tau \subseteq R$ relative to $\operatorname{aff}(R)$ is $k-k'$. Since θ is generic, the essentialization of the collection $\{H_R(\ell, j)\}$ is generic (Definition 3.1). Hence, by Lemma 3.4, τ lies in exactly $(k-k')$ of these hyperplanes. Altogether, τ is contained in exactly $k' + (k-k') = k$ bent hyperplanes, as claimed. □

See Figure 2c for an illustration of a parameter that is not supertransversal. In supertransversal networks one can assign to each facet in the canonical polyhedral complex a unique bent hyperplane, which allows one to compute the tropical weights directly from the parameters.

Proposition 3.6. *Let θ be supertransversal. Let $\sigma \in \mathcal{C}_\theta^{d-1}$ be a facet, and let $B_{\ell,i}$ be the unique bent hyperplane containing σ. For each layer $k \in [L]$, let $S_k \subseteq [n_k]$ denote the set of strictly active neurons on the relative interior of σ. Then the tropical weight $c_\theta(\sigma)$ is given by*

$$c_\theta(\sigma) = \left\| (W^{(1)})^\top D_{S_1} \cdots (W^{(\ell-1)})^\top D_{S_{\ell-1}} (W^{(\ell)})^\top e_i \right\|_2 \; W^{(L+1)} D_{S_L} \cdots W^{(\ell+1)} e_i.$$

Proof. Let $R, Q \in \mathcal{C}_\theta^d$ be the two full-dimensional polyhedra adjacent to σ, and suppose, without loss of generality, that neuron i in layer ℓ is active on R and inactive on Q. Because θ is supertransversal, all other neurons have a fixed sign across σ and its adjacent regions.

On each region, the linear part of f_θ is obtained by multiplying the active weights:

$$A_R = W^{(L+1)} D_{S_L} \cdots W^{(\ell+1)} D_{S_\ell \cup \{i\}} W^{(\ell)} \cdots D_{S_1} W^{(1)},$$

while A_Q is identical except that the index i is excluded from the layer-ℓ diagonal matrix. Across σ, only neuron i changes its state, so the difference in the linear parts is

$$A_R - A_Q = W^{(L+1)} D_{S_L} \cdots W^{(\ell+1)} \operatorname{diag}(e_i) W^{(\ell)} D_{S_{\ell-1}} \cdots W^{(1)}.$$

The normal vector of σ pointing into R is given by the linear part of the i-th preactivation in layer ℓ, normalized to unit length:

$$e_{R/\sigma} = \frac{(W^{(1)})^\top D_{S_1} \cdots (W^{(\ell-1)})^\top D_{S_{\ell-1}} (W^{(\ell)})^\top e_i}{\left\| (W^{(1)})^\top D_{S_1} \cdots (W^{(\ell-1)})^\top D_{S_{\ell-1}} (W^{(\ell)})^\top e_i \right\|_2}.$$

By definition of the tropical weight (Lemma 2.1),

$$c_\theta(\sigma) = (A_R - A_Q) e_{R/\sigma}.$$

We now expand this expression. Using the identity $\mathrm{diag}(e_i)W^{(\ell)}v = e_i\,(e_i^\top W^{(\ell)}v)$, we can factor out the scalar inner product:

$$(A_R - A_Q)e_{R/\sigma} = W^{(L+1)}D_{S_L}\cdots W^{(\ell+1)}e_i\ \underbrace{\left(e_i^\top W^{(\ell)}D_{S_{\ell-1}}\cdots W^{(1)}e_{R/\sigma}\right)}_{\text{scalar norm factor}}.$$

The scalar factor is precisely the norm of the unnormalized linear part:

$$e_i^\top W^{(\ell)}D_{S_{\ell-1}}\cdots W^{(1)}e_{R/\sigma} = \left\|(W^{(1)})^\top D_{S_1}\cdots(W^{(\ell-1)})^\top D_{S_{\ell-1}}(W^{(\ell)})^\top e_i\right\|_2.$$

Substituting this identity yields the desired expression for $c_\theta(\sigma)$. □

Lemma 3.7. *Generic parameters are cancellation-free.*

Proof. Let θ be generic (Definition 3.1). Let $\sigma \in \mathcal{C}_\theta^{d-1}$ be a facet, let $B_{\ell,i}$ be the unique bent hyperplane containing σ, and for each layer $k \in [L]$, let $S_k \subseteq [n_k]$ be the set of active neurons on the relative interior of σ.

By Proposition 3.6, we have

$$c_\theta(\sigma) = \left\|(W^{(1)})^\top D_{S_1}\cdots(W^{(\ell-1)})^\top D_{S_{\ell-1}}(W^{(\ell)})^\top e_i\right\|\ W^{(L+1)}D_{S_L}\cdots W^{(\ell+1)}e_i.$$

Since θ is generic, it satisfies the rank conditions of Definition 3.1. In particular, the first factor (the norm) is zero if and only if there exists a layer $k < \ell$ with $S_k = \emptyset$. Similarly, the forward matrix product is the zero vector if and only if there exists a layer $k > \ell$ with $S_k = \emptyset$.

It follows that $c_\theta(\sigma) = 0$ if and only if there exists a layer $k \in [L]$ such that $S_k = \emptyset$. This is equivalent to $\mathbf{s}_{k,j}(\sigma) = -$ for all $j \in [n_k]$, and hence the parameter is cancellation-free. □

Thus, in the generic regime, codimension in the canonical complex matches the number of bent hyperplanes passing through a face, and the only way a facet can carry zero tropical weight is that all neurons of a later layer are inactive on the corresponding facet.

3.3 Transparency

We now introduce transparency, a simple visibility condition that prevents later layers from annihilating previously created breakpoints. Combined with genericity, transparency will imply the linear region assumption on the relevant domain.

Lemma 3.8. *The network f_θ satisfies LRA on a polyhedron $X \subseteq \mathbb{R}^d$ if and only if $c_\theta(\sigma) \neq 0$ for all facets $\sigma \in \mathcal{C}_\theta^{d-1}$ with $\sigma \cap \mathrm{relint}(X) \neq \emptyset$.*

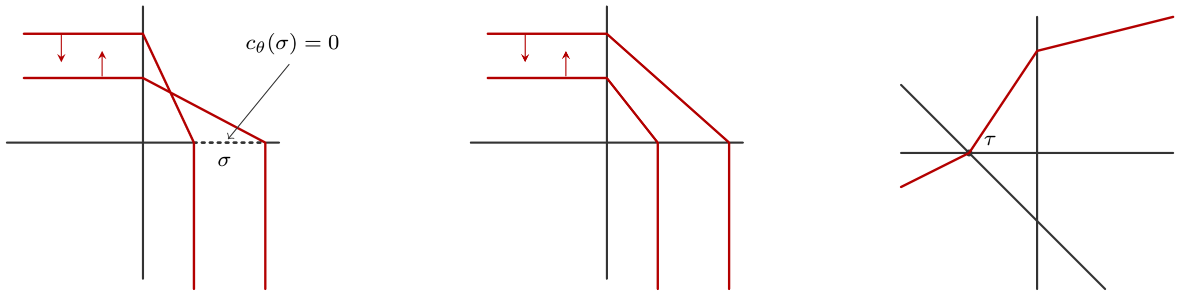
Proof. By definition, f_θ satisfies LRA if and only if no two adjacent maximal polyhedra $P, Q \in \mathcal{C}_\theta^d$ with $(P\cap Q)\cap\mathrm{relint}(X) \neq \emptyset$ share the same affine-linear function. Let $f_\theta|_P(x) = A_Px+b_P$ and $f_\theta|_Q(x) = A_Qx+b_Q$, where A_P and A_Q are the linear parts. Due to the continuity of f_θ across the facet $\sigma = P \cap Q$, these affine-linear functions are identical if and only if their linear parts are equal, $A_P = A_Q$.

According to Lemma 2.1, the weight is $c_\theta(\sigma) = (A_P - A_Q)e_{P/\sigma}$. If $A_P = A_Q$, then $c_\theta(\sigma) = 0$. Conversely, for any CPWL function, the difference between the linear parts on adjacent polyhedra is restricted to the form $A_P - A_Q = ve_{P/\sigma}^\top$ for some $v \in \mathbb{R}^m$. Consequently, $(A_P - A_Q)e_{P/\sigma} = 0$ implies $A_P = A_Q$.

Thus $c_\theta(\sigma) \neq 0$ for every facet σ intersecting $\mathrm{relint}(X)$ if and only if the affine-linear map changes across every facet in the interior of X, which is precisely LRA on X. □

Following Phuong and Lampert (2020), we define transparent ReLU layers as follows.

Definition 3.9. Let $X \subseteq \mathbb{R}^d$. We call a ReLU layer $f_{W,b}\colon \mathbb{R}^d \to \mathbb{R}^m$ *transparent* on X if, for all $x \in X$, there exists an index $i \in [m]$ such that $W_ix + b_i \geq 0$. In words, for every input in X, at least one neuron is active.



(a) A generic parameter that does not satisfy LRA because the second layer is not transparent on σ. In particular, σ would not be contained in the breakpoint complex.

(b) A generic parameter that is transparent and hence satisfies LRA. It also satisfies cTPIC.

(c) A non-supertransversal situation: two first-layer hyperplanes meet at τ, and a second-layer bent hyperplane passes through the same intersection.

Figure 2: Illustration of how transparency implies LRA, and how supertransversality can fail.

Lemma 3.10. *Let $\theta \in \tilde{\Theta}_{\mathcal{A}}$ be generic, and let $X \subseteq \mathbb{R}^d$ be a polyhedron. Suppose that for every $\ell \in \{2, \dots, L\}$, the layer f_{θ_ℓ} is transparent on $a^{(\ell-1)}(X)$. Then θ satisfies LRA on X.*

Proof. Assume that θ does not satisfy LRA on X. Then, by Lemma 3.8, there exists a facet $\sigma \in \mathcal{C}_\theta^{d-1}$ intersecting $\operatorname{relint}(X)$ such that $c_\theta(\sigma) = 0$. Since θ is supertransversal (by Lemma 3.5), there is a unique bent hyperplane $B_{k,j}$ containing σ.

Since θ is cancellation-free, there exists a layer ℓ with $S_\ell(\sigma) = \emptyset$. Since σ lies in the bent hyperplane $B_{k,j}$, no layer $\ell \leq k$ can be entirely inactive on σ; otherwise the preactivation defining $B_{k,j}$ would be locally constant there. Hence $\ell > k$.

This implies that f_{θ_ℓ} is not transparent on $a^{(\ell-1)}(X)$, completing the proof. □

See Figure 2 for an illustration of how transparency implies LRA.

3.4 The Breakpoint Complex

In the case that θ is not transparent, not every piece of every bent hyperplane is necessarily visible as a breakpoint of the final function. The set of breakpoints $B(f_\theta)$ of the function forms the support of a polyhedral complex, $\mathcal{B}_\theta := \{P \in \mathcal{C}_\theta \mid P \subseteq B(f_\theta)\} \subseteq \mathcal{C}_\theta$, which we call the *breakpoint complex*. While the support $B(f_\theta)$ is fully determined by the function f_θ, the complex $\mathcal{B}_\theta$ might in general depend on the specific parameter θ that realizes the function. However, we will prove that for generic parameters, the breakpoint complex $\mathcal{B}_\theta$ only depends on the function f_θ.

Remark 3.11. It follows immediately from Lemma 3.8 that $\mathcal{B}_\theta^{d-1} = \{P \in \mathcal{C}_\theta^{d-1} \mid c_\theta(P) \neq 0\}$.

We aim to use the local geometry of the breakpoint complex to access certain invariants of the neural network. Particularly important are the ridges, which lie in the pairwise intersections of bent hyperplanes.

Definition 3.12. Let $(\mathcal{C}, c)$ be a weighted complex in $\mathbb{R}^d$, and let $\tau \in \mathcal{C}^{d-2}$ be a ridge. We say that a facet $\sigma \in \operatorname{star}(\tau)^{d-1}$ is *non-bending at* τ if there exists another facet $\sigma' \in \operatorname{star}(\tau)^{d-1}$ such that $e_{\sigma/\tau} = -e_{\sigma'/\tau}$ and $c(\sigma) = c(\sigma')$. Otherwise, we say that σ is *bending at* τ. We call the ridge τ *bending* if there exists a facet $\sigma \in \operatorname{star}(\tau)^{d-1}$ that is bending at τ, and *non-bending* otherwise.

See Figure 3 for an illustration of bending and non-bending ridges. A one-hidden-layer ReLU network $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^m$ is a CPWL function whose breakpoints form a hyperplane arrangement. In this case, the weight function is constant along these hyperplanes, as shown in the following proposition.

Proposition 3.13. *Let $\theta = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ be the parameter of a one-hidden-layer network, and let $\sigma \in \mathcal{C}_\theta^{d-1}$. Then we have that $c_{f_\theta}(\sigma) = \sum_{i \in I} W^{(2)}_{:,i} \|W^{(1)}_i\|$, where $I = \{i \in [n_1] \mid H_i = \text{aff}(\sigma)\}$.*

Proof. Let $P, Q \in \mathcal{C}_\theta^d$ be the two d-dimensional polyhedra such that $P \cap Q = \sigma$, and let $e_{P/\sigma}$ be the unit normal vector to σ pointing from Q into P. The function f_θ is affine on P and Q with linear parts A_P and A_Q, respectively. For a one-hidden-layer network, the linear part in any region R is given by $A_R = \sum_{k \in S(R)} W^{(2)}_{:,k} W^{(1)}_k$, where $S(R)$ is the set of active neurons in R.

As we cross from Q into P through the facet σ, the only neurons whose activation state can change are those whose preactivations $z_i^{(1)}(x) = \langle W_i^{(1)}, x\rangle + b_i^{(1)}$ vanish on $\text{aff}(\sigma)$. This is precisely the set I. Thus, the difference in the linear parts is:

$$A_P - A_Q = \sum_{i \in I \cap S(P)} W^{(2)}_{:,i} W^{(1)}_i - \sum_{i \in I \cap S(Q)} W^{(2)}_{:,i} W^{(1)}_i.$$

By definition, $c_{f_\theta}(\sigma) = (A_P - A_Q)e_{P/\sigma}$. For any $i \in I$, the vector $W_i^{(1)}$ is normal to σ, so $W_i^{(1)} = \lambda_i e_{P/\sigma}$ for some $\lambda_i \in \mathbb{R}$. Note that $W_i^{(1)} e_{P/\sigma} = \lambda_i$ and hence $\|W_i^{(1)}\|_2 = |\lambda_i|$. There are two cases for each $i \in I$:

1. If $\lambda_i > 0$, the preactivation $z_i^{(1)}$ increases in the direction of $e_{P/\sigma}$. Thus, $i \in S(P)$ and $i \notin S(Q)$. The contribution to $(A_P - A_Q)e_{P/\sigma}$ is $W^{(2)}_{:,i}\lambda_i = W^{(2)}_{:,i}\|W^{(1)}_i\|$.
2. If $\lambda_i < 0$, the preactivation $z_i^{(1)}$ decreases in the direction of $e_{P/\sigma}$. Thus, $i \notin S(P)$ and $i \in S(Q)$. The contribution to $(A_P - A_Q)e_{P/\sigma}$ is $-W^{(2)}_{:,i}\lambda_i = W^{(2)}_{:,i}|\lambda_i| = W^{(2)}_{:,i}\|W^{(1)}_i\|$.

Summing over all $i \in I$, we obtain $c_{f_\theta}(\sigma) = \sum_{i \in I} W^{(2)}_{:,i}\|W^{(1)}_i\|$. ☐

See also Figure 3c for an illustration.

Lemma 3.14. *If $\tau \in \mathcal{B}_\theta$ lies only on bent hyperplanes from a single layer $\ell \in [L]$, i.e., $\tau \nsubseteq B_{k,j}$ for all neurons (k, j) with $k \neq \ell$, then τ is non-bending.*

Proof. Let $P \in \mathcal{C}^d_{\theta,\ell-1}$ be the maximal polyhedron containing τ. Then $a^{(\ell-1)}$ is affine linear on P, and consequently the breakpoints of $a^{(\ell)}$ on P form a hyperplane arrangement. Thus, each facet in $\text{star}(\tau)^{d-1}$ has an opposite facet. By Proposition 3.13, opposite facets have the same weights, and hence τ is non-bending. ☐

The contraposition of Lemma 3.14 implies that any bending ridge lies in the intersection of bent hyperplanes associated with neurons from different layers. The converse of Lemma 3.14 does not hold in general.

Definition 3.15. We call θ *honest* if the converse of Lemma 3.14 holds, i.e., if $\tau \in \mathcal{B}_\theta$ is non-bending, then it lies only on bent hyperplanes from a single layer.

Lemma 3.16. *Generic parameters are honest.*

Proof. Let θ be generic, and let $\tau \in \mathcal{C}_\theta^{d-2}$ be a ridge. Since θ is generic, τ is contained in precisely two bent hyperplanes. Suppose that these two bent hyperplanes are $B_{k,i}$ and $B_{\ell,j}$, with $k < \ell$. We show that τ is bending.

By supertransversality, there exists an open neighborhood U of $\text{relint}(\tau)$ such that no other bent hyperplanes intersect U. The bent hyperplane $B_{k,i}$ subdivides U into two regions, U^+ and U^-, on which neuron (k, i) is active and inactive, respectively.

Let a^+ and a^- be the local gradients of the preactivation $z_j^{(\ell)}$ on the two sides of $B_{k,i}$. Let n be the local normal of $B_{k,i}$. Since crossing $B_{k,i}$ only changes the activation state of neuron (k, i), the gradient jump of

$z_j^{(\ell)}$ across $B_{k,i}$ has the form $a^+ - a^- = \gamma n$ for some scalar γ. By the rank condition in Definition 3.1, this scalar is nonzero.

We claim that a^+ and a^- are not proportional. If $a^+ = \lambda a^-$, then $\gamma n = (\lambda - 1)a^-$. Since $\gamma \neq 0$, this would imply that n is proportional to a^-. But n and a^- are the local normals of the two bent hyperplanes $B_{k,i}$ and $B_{\ell,j}$ meeting at the codimension-two ridge τ. By genericity, these bent hyperplanes meet transversely there, so their local normals are not proportional. This contradiction shows that a^+ and a^- are not proportional.

Thus the two facets of $B_{\ell,j}$ adjacent to τ do not have opposite relative normal directions. Hence τ is bending. Therefore any non-bending ridge can only lie on bent hyperplanes from a single layer, which is precisely honesty. □

Next, we prove the following for generic parameters.

Lemma 3.17. *Let θ be supertransversal and cancellation-free, and let $\tau \in \mathcal{B}_\theta^{d-2}$. Then there are either three or four facets adjacent to τ: $\#\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} \in \{3, 4\}$.*

Proof. Since $\mathcal{B}_\theta$ is a subcomplex of $\mathcal{C}_\theta$, we have $\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} \subseteq \operatorname{star}_{\mathcal{C}_\theta}(\tau)^{d-1}$. Because θ is supertransversal and τ has codimension 2, τ is contained in exactly two bent hyperplanes. The local arrangement of these two hyperplanes is combinatorially equivalent to the intersection of two transverse affine hyperplanes. Hence $\operatorname{star}_{\mathcal{C}_\theta}(\tau)^{d-1}$ consists of exactly four facets, and therefore $\#\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} \leq 4$.

We show that at least three facets remain. Suppose, for contradiction, that there exists a layer ℓ' such that $\mathbf{s}_{\ell',j}(\tau) = -$ for all $j \in [n_{\ell'}]$. Then f_θ is locally constant in a neighborhood of $\operatorname{relint}(\tau)$, contradicting the assumption that $\tau \in \mathcal{B}_\theta$. Thus, in every layer ℓ', there exists at least one neuron $j \in [n_{\ell'}]$ such that $\mathbf{s}_{\ell',j}(\tau) \neq -$.

Let $B_{\ell',i'}$ and $B_{\ell,i}$ be the two bent hyperplanes containing τ and assume without loss of generality that $\ell' \leq \ell$. Since θ is supertransversal, no bent hyperplanes from layers $k \notin \{\ell', \ell\}$ intersect $\operatorname{relint}(\tau)$. Consequently, for each layer $k \notin \{\ell', \ell\}$, there exists a neuron $j \in [n_k]$ such that $\mathbf{s}_{k,j}(\tau) = +$, and hence, for every facet $\sigma \in \operatorname{star}_{\mathcal{C}_\theta}(\tau)^{d-1}$, we also have that $\mathbf{s}_{k,j}(\sigma) = +$.

Now, let $P, Q \in \mathcal{C}_{\theta,\ell-1}^d$ be the regions such that $\operatorname{relint}(\tau) \subseteq \operatorname{relint}(P \cap Q) \subseteq B_{\ell',i'}$. Since $B_{\ell,i}$ intersects the interiors of both P and Q, the preactivation $z_i^{(\ell)}$ cannot be constant on either region. Consider layer $k = \ell'$. If the only neuron active on P corresponds to the bent hyperplane $B_{\ell',i'}$ intersecting τ, then, by supertransversality, there must exist another neuron that is active on Q whose associated hyperplane cannot intersect τ, and hence must be active on τ. In any case, there exists a neuron $j \in [n_{\ell'}]$ that is active on τ, and therefore active on all facets $\sigma \in \operatorname{star}(\tau)^{d-1}$.

Moreover, among the four facets in $\operatorname{star}_{\mathcal{C}_\theta}(\tau)^{d-1}$, there is only one $\sigma \in \operatorname{star}_{\mathcal{C}_\theta}(\tau)^{d-1}$ with $\mathbf{s}_{\ell,j}(\sigma) = -$. Therefore, at most one facet can be removed when passing to $\mathcal{B}_\theta$, which implies $\#\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} \in \{3, 4\}$. □

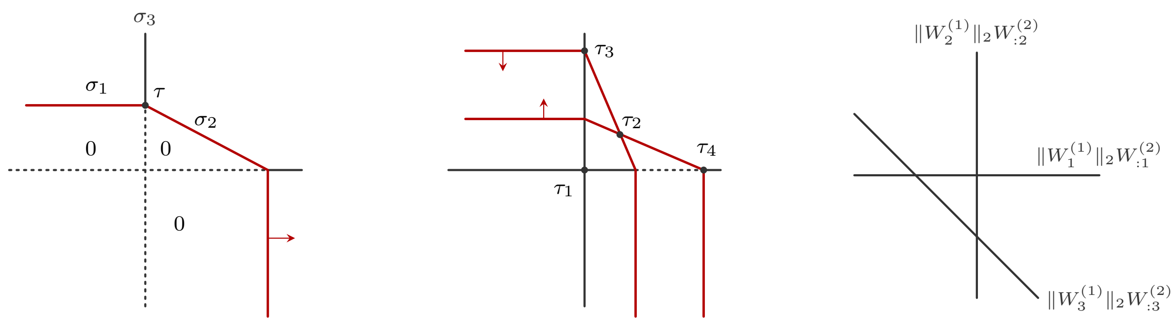
To prove that the breakpoint complex only depends on the function itself and not on the specific parameter representing it, we combine Lemma 3.17 with the following basic lemma.

Lemma 3.18. *Let $\mathcal{B}$ be a polyhedral complex in $\mathbb{R}^d$ that is pure of codimension 1. If for each ridge $\tau \in \mathcal{B}^{d-2}$ one has $\#\operatorname{star}_{\mathcal{B}}(\tau)^{d-1} \geq 3$, then $\mathcal{B}$ is the unique coarsest polyhedral complex on its support.*

Proof. Let $\mathcal{C}$ be a polyhedral complex with $|\mathcal{C}| = |\mathcal{B}|$. We show that $\mathcal{C}$ is a refinement of $\mathcal{B}$.

Assume for contradiction that some facet $P \in \mathcal{C}^{d-1}$ is not contained in any facet of $\mathcal{B}$. Then the intersections $\{\sigma \cap P \mid \sigma \in \mathcal{B}^{d-1},\ \sigma \cap P \neq \emptyset\}$ induce a nontrivial polyhedral subdivision of P. Hence this subdivision has an interior codimension-1 face $F \subset P$. By construction, there are two distinct facets $\sigma_1, \sigma_2 \in \mathcal{B}^{d-1}$ such that $F \subseteq \sigma_1 \cap \sigma_2$. Since $\mathcal{B}$ is pure of codimension 1, the intersection $\sigma_1 \cap \sigma_2$ is a common face of dimension at most $d-2$. Because F has dimension $d-2$, it follows that $\tau := \sigma_1 \cap \sigma_2 \in \mathcal{B}^{d-2}$ and $F \subseteq \tau$.

Now choose $x \in \operatorname{relint}(F)$. Since F is a face of the induced subdivision of P, exactly two cells of that subdivision meet along F near x, namely the traces of σ_1 and σ_2. Therefore, among all facets of $\mathcal{B}$ containing τ, only σ_1 and σ_2 can meet a sufficiently small neighborhood of x inside P. This implies $\#\operatorname{star}_{\mathcal{B}}(\tau)^{d-1} = 2$, contradicting the hypothesis that every ridge of $\mathcal{B}$ is contained in at least three facets.



(a) τ is a bending ridge with 3 adjacent facets. Since the function is constant on regions adjacent to σ_1 and σ_2, they must lie in the bent hyperplane from the later layer.

(b) τ_1 and τ_2 are non-bending ridges, τ_3 is a bending ridge with 4 adjacent facets, and τ_4 is a bending ridge with 3 adjacent facets.

(c) A hyperplane arrangement of a one-hidden-layer network. The weights on the facets are constant along each hyperplane.

Figure 3: Illustration of bending and non-bending ridges.

Hence every facet of $\mathcal{C}$ is contained in a facet of $\mathcal{B}$, so $\mathcal{C}$ is a refinement of $\mathcal{B}$. Thus $\mathcal{B}$ is coarsest. Uniqueness follows because two coarsest complexes on the same support must refine each other. □

Proposition 3.19. *If θ, η are cancellation-free and supertransversal, and satisfy $f_\theta = f_\eta$, then $\mathcal{B}_\theta = \mathcal{B}_\eta$.*

Proof. Since $f_\eta = f_\theta$, the support of $\mathcal{B}_\theta$ equals the support of $\mathcal{B}_\eta$. By Lemma 3.17 and Lemma 3.18, the complexes must be the same. □

Having this at hand, we prove that LRA is a property of the function for generic parameters.

Lemma 3.20. *Let $\theta, \eta \in \widetilde{\Theta}_{\mathcal{A}}$ be generic parameters with $f_\eta = f_\theta$, and let $P \subseteq \mathbb{R}^d$ be a full-dimensional polyhedron. If θ satisfies LRA on P, then η also satisfies LRA on P.*

Proof. Since θ and η are generic, they are supertransversal and cancellation-free by Lemmas 3.5 and 3.7. Since $f_\eta = f_\theta$, their breakpoint complexes coincide by Proposition 3.19: $\mathcal{B}_\eta = \mathcal{B}_\theta$.

Assume that θ satisfies LRA on P. Then every codimension-one face of the canonical complex of θ meeting $\operatorname{relint}(P)$ is visible, and hence every ridge $\tau \in \mathcal{B}_\theta^{d-2}$ with $\tau \cap \operatorname{relint}(P) \neq \emptyset$ is adjacent to all four local facets of the canonical complex. By Lemma 3.17, the only possible numbers of adjacent breakpoint facets are 3 and 4, so for θ every such ridge has exactly 4 adjacent facets in $\mathcal{B}_\theta$.

Since $\mathcal{B}_\eta = \mathcal{B}_\theta$, the same is true for η. Therefore no codimension-one face of the canonical complex of η meeting $\operatorname{relint}(P)$ is lost when passing to the breakpoint complex, and hence η satisfies LRA on P. □

3.5 The Dependency Graph

We now show how the local geometry of the breakpoint complex encodes layer precedence for generic parameters. The key observation is that bending ridges distinguish earlier from later bent hyperplanes, and therefore allow us to organize visible nonlinearities into a directed dependency graph.

The following lemma demonstrates that, at a bending ridge, we can distinguish which adjacent facets originate from the earlier layer and which from the later one. See also Figure 3b for an illustration.

Lemma 3.21. *Let θ be generic. Let $\tau \in \mathcal{B}_\theta^{d-2}$ be a bending ridge, and denote the unique bent hyperplanes containing τ as $H \in \mathcal{H}_k(\theta)$, $B \in \mathcal{H}_\ell(\theta)$ with $k < \ell$. Let $R \in \mathcal{C}_{\theta,k-1}^d$ be the region such that $\operatorname{relint}(\tau) \subseteq \operatorname{relint}(R)$. Then:*

1. *If $\#\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} = 4$, then there are exactly two facets $\sigma_1, \sigma_2 \in \operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1}$ such that $e_{\sigma_1/\tau} = -e_{\sigma_2/\tau}$. In this case, $H \cap R = \operatorname{aff}(\sigma_1) \cap R$.*

2. *If* $\#\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} = 3$*, then there is a unique facet* $\sigma \in \operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1}$ *that is not adjacent to a region* $P \in \operatorname{star}_{\mathcal{C}_\theta}(\tau)^d$ *on which* $f_\theta|_P$ *is constant. In this case,* $H \cap R = \operatorname{aff}(\sigma) \cap R$*.*

Proof. Case 1: $\#\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} = 4$. In this case, all four facets in $\operatorname{star}_{\mathcal{C}_\theta}(\tau)^{d-1}$ are contained in $\mathcal{B}_\theta$. Exactly two of these facets are contained in H. Since $H \cap R$ is the zero-locus of an affine-linear function, both facets have the same affine span. Thus there exist exactly two facets $\sigma_1, \sigma_2 \in \operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1}$ such that $\operatorname{aff}(\sigma_1) = \operatorname{aff}(\sigma_2)$ (and hence $e_{\sigma_1/\tau} = -e_{\sigma_2/\tau}$). For these facets, we have $H \cap R = \operatorname{aff}(\sigma_1) \cap R$. The remaining two facets cannot have the same affine span, because τ is bending.

Case 2: $\#\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1} = 3$. In this case, exactly one of the two facets contained in H is missing from $\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1}$. Let σ_1 denote the missing facet and σ_2 the remaining facet with $\sigma_2 \subseteq H$.

By cancellation-freeness, $\sigma_1 \notin \mathcal{B}_\theta$ implies that there exists a layer in which all neurons are inactive on σ_1. Hence f_θ is constant on σ_1 and on the two adjacent regions $P_1, Q_1 \in \operatorname{star}_{\mathcal{C}_\theta}(\tau)^d$. These two regions are also adjacent to the two facets in $\operatorname{star}_{\mathcal{C}_\theta}(\tau)^{d-1}$ that are contained in B. Let $P_2, Q_2 \in \operatorname{star}_{\mathcal{C}_\theta}(\tau)^d$ be the two regions adjacent to σ_2, and denote by $\sigma_P = P_1 \cap P_2$ and $\sigma_Q = Q_1 \cap Q_2$ the corresponding facets contained in B. If f_θ were constant on either P_2 or Q_2, then σ_P or σ_Q would also be absent from $\mathcal{B}_\theta$, contradicting the assumption that exactly one facet is missing. Therefore, f_θ is non-constant on both regions adjacent to σ_2.

It follows that among the three facets in $\operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1}$ there is exactly one facet that is not adjacent to any region on which f_θ is constant. This facet must be σ_2, and since $\sigma_2 \subseteq H$, we obtain $H \cap R = \operatorname{aff}(\sigma_2) \cap R$. See also Figure 3a for an illustration. □

By leveraging the geometric asymmetry established in Lemma 3.21, we reconstruct a directed graph from the breakpoints of f_θ that reflects the relative layer order of the associated neurons.

Definition 3.22 (Candidate bent hyperplanes)**.** Let θ be generic and let $\mathcal{B}_\theta$ be the breakpoint complex. We define an equivalence relation on the facets $\mathcal{B}_\theta^{d-1}$ as follows.

Let $\tau \in \mathcal{B}_\theta^{d-2}$ be a ridge and let $\sigma, \sigma' \in \operatorname{star}_{\mathcal{B}_\theta}(\tau)^{d-1}$.

- If τ is non-bending, we declare $\sigma \sim \sigma'$ whenever $e_{\sigma/\tau} = -e_{\sigma'/\tau}$ and $c_\theta(\sigma) = c_\theta(\sigma')$. Thus opposite facets with the same weight are continuations of the same candidate bent hyperplane.
- If τ is bending, we declare $\sigma \sim \sigma'$ whenever the local description of Lemma 3.21 identifies them as belonging to the same one of the two bent hyperplanes through τ.

A *candidate bent hyperplane* is an equivalence class of facets under the transitive closure of this relation. The set of candidate bent hyperplanes is denoted by V_{f_θ}.

Definition 3.23. We define the *dependency graph* (V_f, E_f) as follows. The vertex set V_f consists of the candidate bent hyperplanes. The edge set E_f consists of all pairs $(v, w) \in V_f \times V_f$ for which there exists a bending ridge τ where w "bends" along v as in Lemma 3.21.

Remark 3.24. The definition of the dependency graph is a generalization of the definition given in the work of Phuong and Lampert (2020) for transparent networks.

The following proposition establishes that any generic parameter realizing f_θ must respect the partial order induced by this graph.

Proposition 3.25. *Let* θ *be a generic parameter of some architecture, and let* $(V_{f_\theta}, E_{f_\theta})$ *be the corresponding dependency graph. Let* $\mathcal{A} = (n_0, \ldots, n_L, m)$ *be a (potentially different) architecture and* $V_\mathcal{A} = \{(i, \ell) \mid \ell \in [L],\, i \in [n_\ell]\}$ *be the set of hidden neurons. Then, for a generic parameter* $\eta \in \tilde{\Theta}_\mathcal{A}$ *with* $f_\eta = f_\theta$*, there is a map* $\phi_\eta \colon V_{f_\theta} \to V_\mathcal{A}$ *such that for every* $(u, v) \in E_{f_\theta}$*, if* $(i, \ell) = \phi_\eta(u)$ *and* $(j, k) = \phi_\eta(v)$*, then* $\ell < k$*.*

Proof. Since θ and η are generic, it is supertransversal and cancellation-free by Lemmas 3.5 and 3.7, and honest by Lemma 3.16. Moreover, since $f_\eta = f_\theta$, the breakpoint complex is determined by the function, and hence $\mathcal{B}_\eta = \mathcal{B}_\theta$ by Proposition 3.19. Thus the local weighted geometry of the breakpoint complex is the same for the realizations θ and η.

We first define the map ϕ_η. Let $u \in V_{f_\theta}$ be a candidate bent hyperplane, i.e., an equivalence class of facets in $\mathcal{B}_\theta^{d-1}$ under the transitive closure of direct equivalence. Choose any facet $\sigma \in u$. Since σ is a facet of the breakpoint complex of f_η, it is contained in the bent hyperplane of some neuron of the realization η. We define $\phi_\eta(u)$ to be such a neuron.

We must check that this is well defined on the whole class u. Suppose $\sigma, \sigma' \in u$ are directly equivalent. Then, by definition, there exists a ridge $\tau \in \mathcal{B}_\theta^{d-2}$ with $\sigma, \sigma' \in \operatorname{star}(\tau)$ such that σ and σ' come from the same layer according to Lemma 3.21. Because the weighted breakpoint complex depends only on the realized function, the same local configuration occurs for η. Hence the two facets σ and σ' must also belong to the same local bent-hyperplane piece in the realization η. Passing to the transitive closure, all facets in the class u are assigned consistently, and therefore ϕ_η is well defined.

Now let $(u, v) \in E_{f_\theta}$. By definition of the dependency graph, there exists a bending ridge $\tau \in \mathcal{B}_\theta^{d-2}$ such that the candidate bent hyperplane v bends along u. Let $\phi_\eta(u) = (i, \ell), \phi_\eta(v) = (j, k)$. We claim that $\ell < k$.

Since θ is honest, the ridge τ lies in bent hyperplanes coming from two different layers. By Lemma 3.21, the local geometry at τ determines which adjacent facets belong to the earlier bent hyperplane and which belong to the later one. Because $\mathcal{B}_\eta = \mathcal{B}_\theta$ as weighted complexes, the same distinction must hold for the realization η. Thus the neuron of η corresponding to the class u must lie in a strictly earlier layer than the neuron corresponding to the class v. Hence $\ell < k$.

Therefore, for every edge $(u, v) \in E_{f_\theta}$, the map ϕ_η sends u and v to neurons whose layer indices satisfy $\ell < k$. This proves the claim. □

3.6 A Sufficient Condition for Identifiability Among Generic Parameters

We say that a parameter $\theta \in \tilde{\Theta}_{\mathcal{A}}$ is *identifiable among generic parameters* if $\eta \in \tilde{\Theta}_{\mathcal{A}}$ with $f_\eta = f_\theta$ implies $\eta \sim \theta$. In this subsection, we derive a sufficient criterion for identifiability among generic parameters. The argument has three steps. First, we isolate a geometric regime in which the reverse-engineering procedure of Rolnick and Kording (2020) applies. Second, we show with the breakpoint complex and the dependency graph that the relevant visibility and transversality conditions are properties of function on the generic locus. Third, we use the rank conditions built into our notion of genericity to remove a remaining sign ambiguity in the last hidden layer resulting from the reverse-engineering procedure.

Definition 3.26. Let $P \subseteq \mathbb{R}^d$ be a polyhedron, and let $B_1, B_2 \subseteq \mathbb{R}^d$ be two piecewise linear hypersurfaces. We say B_1 and B_2 *intersect transversely* in P if $B_1 \cap B_2 \cap \operatorname{relint}(P) \neq \emptyset$ and $B_1 \cap B_2 \cap P$ is pure of dimension $\dim(P) - 2$. A ReLU network $f : \mathbb{R}^d \to \mathbb{R}^m$ satisfies *TPIC* (Transverse Pairwise Intersection Condition) on P if, for every pair of adjacent neurons $(j, \ell - 1)$ and (i, ℓ), the corresponding bent hyperplanes intersect transversely in P.

The condition TPIC is the natural parameter-level transversality condition used in the literature. However, for our purposes it is useful to strengthen it to the following cTPIC, which keeps track of connected visible pieces of the bent hyperplanes on the working polytope.

Definition 3.27. Let $\theta \in \Theta_{\mathcal{A}}$ and let $P \subseteq \mathbb{R}^d$ be a polyhedron. We say that f_θ satisfies the *connected transverse pairwise intersection condition* (cTPIC) on P if, for every hidden neuron (j, ℓ), there exists a connected piecewise linear hypersurface $C_{j,\ell} \subseteq B_{j,\ell}(\theta)$ such that for every adjacent-layer pair of neurons (j, ℓ) and $(i, \ell + 1)$, the sets $C_{j,\ell}$ and $C_{i,\ell+1}$ intersect transversely in P.

See Figure 2 examples of networks that satisfy cTPIC on the entire domain. The strengthening from TPIC to cTPIC is tailored to our later use of the dependency graph: the connected pieces $C_{j,\ell}$ allow us to identify, from the breakpoint complex, the candidate bent hyperplanes corresponding to individual neurons. Before turning to this transfer mechanism, we recall the reverse-engineering theorem on the restricted class of parameters where it applies.

To isolate the geometric regime in which the reverse-engineering algorithm of Rolnick and Kording (2020) applies, we bundle these assumptions into the following class of parameters. Let $\Theta_{\mathcal{A}}^{\text{good}}$ be the set of parameters in $\widetilde{\Theta}$ that satisfy cTPIC and LRA on some full-dimensional polytope. The reverse-engineering argument of Rolnick and Kording (2020) should be interpreted as a reconstruction theorem inside the class $\Theta_{\mathcal{A}}^{\mathrm{good}}$. In particular, it yields uniqueness only among parameters in this restricted class, not yet among all generic parameters or all parameters; see Appendix A.1. The resulting reconstruction is unique up to permutation and positive scaling in the first $L-1$ hidden layers, with a remaining neuronwise sign ambiguity in the final hidden layer.

Theorem 3.28 (Reverse-engineering up to sign, Rolnick and Kording, 2020)**.** *Let $\theta, \eta \in \Theta_{\mathcal{A}}^{\mathrm{good}}$ satisfy $f_\eta = f_\theta$. Then, up to permutation and positive scaling, the first $L-1$ hidden layers of η and θ coincide, and the L-th hidden layer coincides up to permutation, positive scaling, and neuronwise sign.*

Before addressing this remaining sign ambiguity, we explain how our breakpoint-complex framework transfers the properties cTPIC and LRA across generic realizations of the same function. This is the step that upgrades the reverse-engineering theorem from a uniqueness statement inside the restricted class $\Theta_{\mathcal{A}}^{\mathrm{good}}$ to a statement about all generic realizations.

Lemma 3.29. *Let $\theta \in \widetilde{\Theta}_{\mathcal{A}}$ be generic and suppose that f_θ satisfies cTPIC on a full-dimensional polytope P. Then the dependency graph $(V_{f_\theta}, E_{f_\theta})$ contains a layered subgraph whose vertices correspond to the hidden neurons of $\mathcal{A}$, and in which every pair of adjacent layers is fully connected.*

Proof. For each hidden neuron (j, ℓ), cTPIC provides a connected piecewise linear hypersurface $C_{j,\ell} \subseteq B_{j,\ell}(\theta)$ of codimension one such that the ones from adjacent neurons intersect transversely in P. Since θ is generic, every bending ridge of the breakpoint complex lies in exactly two bent hyperplanes, and by Lemma 3.21 the local weighted geometry determines which adjacent facets belong to the earlier and later bent hyperplanes. Because each $C_{j,\ell}$ is connected, all its facets belong to a single candidate bent hyperplane, and hence define a vertex of the dependency graph.

Now let (j, ℓ) and $(i, \ell+1)$ be neurons in adjacent layers. By cTPIC, the sets $C_{j,\ell}$ and $C_{i,\ell+1}$ intersect transversely in P. Their intersection contains a bending ridge of the breakpoint complex, and by construction this ridge yields an edge in the dependency graph from the candidate bent hyperplane corresponding to (j, ℓ) to the one corresponding to $(i, \ell+1)$. Since this holds for every adjacent-layer pair, the resulting vertices form a layered subgraph with all adjacent layers fully connected. □

The first lemma shows that cTPIC produces the layered combinatorics required by the dependency graph. Conversely, the next lemma shows that this layered subgraph structure is sufficient to recover cTPIC on the generic locus.

Lemma 3.30. *Let $\theta \in \widetilde{\Theta}_{\mathcal{A}}$ be generic. If the dependency graph $(V_{f_\theta}, E_{f_\theta})$ contains a layered subgraph corresponding to the hidden architecture, with all adjacent layers fully connected, then f_θ satisfies cTPIC.*

Proof. Let (j, ℓ) and $(i, \ell+1)$ be neurons in adjacent layers. By assumption, the dependency graph contains an edge between the corresponding vertices. By definition of the dependency graph, this edge arises from a bending ridge τ of the breakpoint complex at which the later candidate bent hyperplane bends along the earlier one. Since θ is generic, Lemma 3.21 implies that τ lies in the intersection of the two corresponding bent hyperplanes and that this intersection has codimension 2. Hence the two bent hyperplanes intersect transversely. For each vertex of the layered subgraph, take the union of the facets belonging to the corresponding candidate bent hyperplane. By construction this union is connected, and the preceding argument shows that adjacent-layer representatives intersect transversely. Hence these sets witness cTPIC. □

The example in the work of (Ramakrishnan, 2026) shows that the remaining ambiguity in Theorem 3.28 is genuine: even under cTPIC and LRA, the last hidden layer need not be identifiable from the bent hyperplanes alone. In fact, the phenomenon already appears in the simplest setting of a network with a single hidden layer, where no deeper hidden layer is available to resolve the orientation of the breakpoint hyperplanes. In that case, the ambiguity can be interpreted as arising from a cancellation among the rank-one matrices $W^{(2)}_{:,j} W^{(1)}_{j,:}$. More generally, in our framework such examples arise from a nontrivial cancellation in the output contribution of the final hidden layer: a sign flip of a nonempty subset $J \subseteq [n_L]$ can remain invisible in the realized function only if the corresponding masked products $W^{(L+1)}D_J W^{(L)} D_{S_{L-1}(R)} W^{(L-1)} \cdots D_{S_1(R)} W^{(1)}$ vanish on the relevant full-dimensional regions $R \in \mathcal{C}_{\theta,L-1}$.

The definition of genericity excludes precisely this type of cancellation. On the other hand, Lemmas 3.20, 3.29 and 3.30 show that the assumptions needed for reverse engineering are functionally determined on the generic locus. Combining these two ingredients upgrades Theorem 3.28 from a reconstruction statement inside $\Theta_{\mathcal{A}}^{\text{good}}$ to identifiability among generic parameters.

Theorem 3.31. *Let $\theta \in \tilde{\Theta}_{\mathcal{A}}$ be generic. If θ satisfies cTPIC and LRA on a full-dimensional polyhedron $P \subseteq \mathbb{R}^d$, then θ is identifiable among generic parameters.*

Proof. Let $\eta \in \widetilde{\Theta}_{\mathcal{A}}$ satisfy $f_\eta = f_\theta$. Since θ is generic and satisfies cTPIC and LRA, we have $\theta \in \Theta_{\mathcal{A}}^{\text{good}}$.

By Lemma 3.20, the parameter η also satisfies LRA. Moreover, by the functional invariance of the dependency graph on the generic locus and Lemmas 3.29 and 3.30, the parameter η also satisfies cTPIC. Therefore $\eta \in \Theta_{\mathcal{A}}^{\text{good}}$.

By Theorem 3.28, after applying permutation and positive scaling symmetries to η, we may assume that $(W^{(\ell,\eta)}, b^{(\ell,\eta)}) = (W^{(\ell,\theta)}, b^{(\ell,\theta)})$ for all $\ell \in [L-1]$, and that for each $j \in [n_L]$ there is a sign $s_j \in \{\pm 1\}$ such that $z_j^{(L,\eta)} = s_j\, z_j^{(L,\theta)}$.

We first determine the output layer. Let $\sigma \in \mathcal{C}_\theta^{d-1}(P)$ be a facet contained in the bent hyperplane of a neuron (j, L) in the last hidden layer. Since θ is generic, it is supertransversal and cancellation-free by Lemmas 3.5 and 3.7. Hence Proposition 3.6 applies and gives $c_\theta(\sigma) = \left\|(W^{(1,\theta)})^\top D_{S_1(\sigma)} \cdots (W^{(L,\theta)})^\top e_j\right\|_2 \; W^{(L+1,\theta)} e_j$. The reverse-engineering procedure determines the local hyperplane of the bent hyperplane containing σ, and therefore determines the norm $\left\|(W^{(1,\theta)})^\top D_{S_1(\sigma)} \cdots (W^{(L,\theta)})^\top e_j\right\|_2$ from the function. Since $c_\theta(\sigma)$ is also determined by the function, the column $W^{(L+1,\theta)} e_j$ is uniquely determined. Repeating this for all neurons in the last hidden layer recovers $W^{(L+1,\theta)}$ uniquely. The output bias $b^{(L+1,\theta)}$ is then determined by evaluating the affine-linear piece of the function on any full-dimensional region. Thus, after replacing η by an equivalent parameter if necessary, we may assume $(W^{(L+1,\eta)}, b^{(L+1,\eta)}) = (W^{(L+1,\theta)}, b^{(L+1,\theta)})$.

It remains to show that $s_j = +1$ for all $j \in [n_L]$. Suppose not, and let $J \coloneqq \{j \in [n_L] \mid s_j = -1\}$. Then $J \neq \emptyset$. Choose $j \in J$. Since the bent hyperplane $B_{j,L}(\theta)$ is visible since θ satisfies LRA, there exists a full-dimensional polyhedron $R \in \mathcal{C}_{\theta,L-1}^d(P)$ such that $B_{j,L}(\theta) \cap \operatorname{relint}(R) \neq \emptyset$. Equivalently, the restricted preactivation $z_j^{(L,\theta)}|_R$ is non-constant. We claim that this implies $S_k(R) \neq \emptyset$ for all $k \in [L-1]$. Indeed, if $S_k(R) = \emptyset$ for some $k \in [L-1]$, then $a^{(k,\theta)}$ is identically zero on R. Consequently, $z^{(k+1,\theta)} = W^{(k+1,\theta)} a^{(k,\theta)} + b^{(k+1,\theta)} = b^{(k+1,\theta)}$ is constant on R, and therefore $a^{(k+1,\theta)}$ is constant on R as well. Iterating this argument shows that $z^{(\ell',\theta)}$ is constant on R for all $\ell' \geq k+1$, in particular that $z_j^{(L,\theta)}|_R$ is constant, a contradiction.

On R, the linear part of the last hidden preactivation is $A_R \coloneqq W^{(L,\theta)} D_{S_{L-1}(R)} W^{(L-1,\theta)} \cdots D_{S_1(R)} W^{(1,\theta)}$. For each $j \in J$, the contribution of the j-th neuron in the last hidden layer changes from $W_{:,j}^{(L+1,\theta)}[z_j^{(L,\theta)}]_+$ to $W_{:,j}^{(L+1,\theta)}[-z_j^{(L,\theta)}]_+$. Using the identity $[-u]_+ = [u]_+ - u$, the difference in the contribution of neuron j is $-W_{:,j}^{(L+1,\theta)} z_j^{(L,\theta)}$. Hence the difference $f_\eta - f_\theta$ on R is affine-linear with linear part $-W^{(L+1,\theta)} D_J A_R = -W^{(L+1,\theta)} D_J W^{(L,\theta)} D_{S_{L-1}(R)} W^{(L-1,\theta)} \cdots D_{S_1(R)} W^{(1,\theta)}$. Since $f_\eta = f_\theta$, this linear part must vanish:

$$W^{(L+1,\theta)} D_J W^{(L,\theta)} D_{S_{L-1}(R)} W^{(L-1,\theta)} \cdots D_{S_1(R)} W^{(1,\theta)} = 0.$$

By Condition (2) in Definition 3.1, the above matrix has rank $\min\big(d,\ n_{L+1},\ |J|,\ |S_1(R)|, \ldots, |S_{L-1}(R)|\big)$. Since $J \neq \emptyset$ and $S_k(R) \neq \emptyset$ for all $k \in [L-1]$, this minimum is at least 1. Hence the matrix is nonzero, a contradiction.

Therefore $J = \emptyset$, so $s_j = +1$ for all $j \in [n_L]$. Thus the final hidden layer is also uniquely determined up to permutation and positive scaling. We conclude that every generic parameter η with $f_\eta = f_\theta$ is equivalent to θ, and hence θ is identifiable among generic parameters. ☐

4 Identifiable Parameters in Most Architectures

In this section, we show that for every architecture with layers of widths at least 2, there exists a nonempty open subset of identifiable parameters. We first construct an open set on which functions determine parameters among generic parameters, and then upgrade this to identifiability among all parameters by excluding the lower-dimensional set admitting nongeneric parameters in the fiber.

Our strategy for the first part is to construct a generic parameter satisfying cTPIC and LRA, and then use the openness of these conditions. Our construction is inspired by the inductive bent-hyperplane approach of Grigsby et al. (2023), but refines it by using bounded polytopes and slab layers to keep track of which induced nonlinearities remain visible in the final function. This refinement enables us to extend the construction beyond the constraint $n_\ell \geq d$ and to ensure preservation of LRA. In particular, while the inductive construction of Grigsby et al. (2023) enforces transverse intersections at the level of the canonical polyhedral complex, additional care is needed to ensure these intersections remain visible as breakpoints of the function. We discuss this in more detail in Appendix A.2.

4.1 Slab Layers

We aim to construct a sequence of polyhedra in the domains of the intermediate layers on which the two consecutive layers satisfy cTPIC, and then pull this structure back to the input space. The individual layers will be *slab layers* on these polytopes.

Definition 4.1. Let $P \subseteq \mathbb{R}^d$ be a polyhedron. We say that a hyperplane H is *inside* P if it intersects P transversely, that is, if $\operatorname{relint}(P) \cap H \neq \emptyset$ and $\dim(P \cap H) = \dim(P) - 1$. A hyperplane arrangement $\mathcal{H} \subseteq \mathbb{R}^d$ is *inside* P if every $H \in \mathcal{H}$ is inside P. We say that a ReLU layer $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^n$ is a *slab layer* on P if

1. H is inside P for all $H \in \mathcal{H}_\theta$, and
2. $H_1 \cap H_2 \cap P = \emptyset$ for all $H_1, H_2 \in \mathcal{H}_\theta$.

Definition 4.2. Let $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^n$ be a slab layer on P. Let $H_1, \ldots, H_n$ be the hyperplanes of the slab layer, ordered such that they partition P into regions $R_0, \ldots, R_n \in \mathcal{C}_\theta(P)$ with $R_{k-1} \cap R_k = H_k \cap P$. Then we call the slab layer *oriented* if $S(R_i) = \{1, \ldots, i\} \triangle \{n\}$, where $\triangle$ denotes the symmetric difference.

See Figure 4a for an illustration of an oriented slab layer that partitions a polytope into regions. Since our goal is to construct an open set, we need to ensure the notions we introduce are open, meaning that they are preserved under sufficiently small perturbations.

Definition 4.3. Let $H = \{x \in \mathbb{R}^d \mid \langle w, x \rangle + b = 0\}$ be a hyperplane and let $\varepsilon > 0$. We say that a hyperplane $H' = \{x \in \mathbb{R}^d \mid \langle w', x \rangle + b' = 0\}$ is ε-close to H if

$$\min \left\{ \left\| \frac{(w', b')}{\|w'\|} - \frac{(w, b)}{\|w\|} \right\|_\infty, \left\| \frac{(w', b')}{\|w'\|} - \frac{(-w, -b)}{\|w\|} \right\|_\infty \right\} < \varepsilon.$$

A hyperplane arrangement $\mathcal{H}$ is ε-close to H if all its hyperplanes are ε-close to H.

Remark 4.4. If a hyperplane H is inside a polyhedron P (Definition 4.1), then there is an $\varepsilon > 0$ such that all hyperplanes H' that are ε-close to H are also inside P.

Next, given a polytope P inside a bounded set X and a hyperplane inside P, we construct an oriented slab layer inside P and transparent on X, by suitably perturbing and orienting the hyperplanes. For the iterative construction, the reader can think of X as the image of the previously constructed layers on a prescribed polytope in the input space.

Lemma 4.5. *Let $X \subseteq \mathbb{R}^d$ be a bounded set, $P \subseteq X$ a polytope, H a hyperplane inside P, $\varepsilon > 0$, and $n \geq 2$. Then there exists a generic oriented slab layer $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^n$ on P that is transparent on X and whose hyperplane arrangement $\mathcal{H}_\theta$ is ε-close to H.*

Proof. Let H be given by $w^\top x + b = 0$. Since H is inside P, it intersects $\mathrm{relint}(P)$ transversely. Because transversality is an open condition, there exists a neighborhood of (w, b) such that all hyperplanes in this neighborhood remain inside P. Choose n distinct values $b_1 > \cdots > b_n$ in an ε neighborhood of b and define perfectly parallel hyperplanes H'_i with weight vectors $\pm w$ such that the activation pattern follows $S(R_i) = \{1, \ldots, i\} \triangle \{n\}$.

In this parallel configuration, neuron n is active on the half-space V_n containing $R_0, \ldots, R_{n-1}$, and neurons $1, \ldots, n-1$ are active on the half-spaces $V_1, \ldots, V_{n-1}$ containing R_n. Since $b_1 > b_n$, the union of the active half-spaces $V_1 \cup V_n$ covers the entire space $\mathbb{R}^d$, ensuring that the layer is transparent. To maintain transparency on X under perturbation, the intersection of the inactive half-spaces $V_1^c \cap V_n^c$ must remain disjoint from X.

Since X and P are bounded sets, there are sufficiently small generic perturbations $\eta_1, \ldots, \eta_n$ of the weight vectors (in particular, $\|\eta_i - w\| < \varepsilon$) such that the intersections $H_i \cap H_j$ of the perturbed hyperplanes lie outside the bounded set X and each H_i remains inside P. In this way, we ensure that for every $x \in X$, either neuron n or neuron 1 is active. In particular, the resulting layer is a generic oriented slab layer on P that is transparent on X and the hyperplane arrangement is ε-close to H. □

4.2 From Slab Layers to cTPIC

To carry out the inductive construction, we need to compose a slab layer with another slab layer in such a way that all bent hyperplanes intersect. The following lemma provides an intermediate step by identifying a neuron with this property for the second layer.

Lemma 4.6. *Let $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^n$ be a generic oriented slab layer on a polyhedron $P \subseteq \mathbb{R}^d$. Then there exist a hyperplane $H' \subseteq \mathbb{R}^n$ and a region $R \in \mathcal{C}_\theta(P)$ such that:*

1. *$f_\theta^{-1}(H') \cap P$ is connected;*
2. *for every $H \in \mathcal{H}_\theta$, the preimage $f_\theta^{-1}(H')$ intersects H transversely in P;*
3. *H' is inside $f_\theta(R)$.*

Proof. Let $H_1, \ldots, H_n$ be the hyperplanes of the slab layer, ordered such that they partition P into regions $R_0, \ldots, R_n$ with $R_{k-1} \cap R_k = H_k \cap P$. For each $k \in [n]$, pick a point $x^{(k)} \in \mathrm{relint}(H_k \cap P)$ and let $y^{(k)} = f_\theta(x^{(k)}) \in \mathbb{R}^n$. The j-th component of $y^{(k)}$ is given by $y_j^{(k)} = [W_j^{(1)} x^{(k)} + b_j^{(1)}]_+$.

By the definition of an oriented slab layer ($S(R_i) = \{1, \ldots, i\} \triangle \{n\}$), the neurons active on the k-th hyperplane are $S(H_k) = S(R_{k-1}) \cap S(R_k)$. Specifically, for $k < n$, $S(H_k) = \{1, \ldots, k-1, n\}$, and for $k = n$, $S(H_n) = \{1, \ldots, n-1\}$. Hence, the matrix $Y = [y^{(1)}, \ldots, y^{(n)}]$ of image points has the following structure:

$$Y = \begin{pmatrix} 0 & y_1^{(2)} & y_1^{(3)} & \ldots & y_1^{(n-1)} & y_1^{(n)} \\ 0 & 0 & y_2^{(3)} & \ldots & y_2^{(n-1)} & y_2^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & y_{n-2}^{(n-1)} & y_{n-2}^{(n)} \\ 0 & 0 & 0 & \ldots & 0 & y_{n-1}^{(n)} \\ y_n^{(1)} & y_n^{(2)} & y_n^{(3)} & \ldots & y_n^{(n-1)} & 0 \end{pmatrix},$$

where all $y_j^{(i)}$ are nonzero. The submatrix formed by the first $n-1$ rows and the last $n-1$ columns is upper-triangular with strictly positive diagonal entries $y_{k-1}^{(k)}$, $k = 2, \ldots, n$. This ensures these $n-1$ columns are linearly independent. Since $y^{(1)}$ is the only column with a non-zero entry in the n-th row and zeros in the first $n-1$ rows, the full set of n columns is linearly independent. In $\mathbb{R}^n$, linear independence of n vectors implies their affine independence, uniquely defining a hyperplane H' passing through these points.

Now, for $m^{(i)} = \frac{1}{2}(y^{(i)} + y^{(i+1)})$, we have that $m^{(i)} \in H'$ and $f_\theta^{-1}(m^{(i)}) \cap \mathrm{relint}(R_i) \neq \emptyset$. Thus $f_\theta^{-1}(H') \cap \mathrm{relint}(R_i) \neq \emptyset$. This implies that $f_\theta^{-1}(H')$ and H_i intersect transversely in P.

Indeed, since H' is the affine hull of $\{y^{(1)}, \dots, y^{(n)}\}$, the midpoint $m^{(i)} = \frac{1}{2}(y^{(i)} + y^{(i+1)})$ is contained in H'. The ReLU layer f_θ is affine-linear on the region R_i bounded by H_i and H_{i+1}. By linearity, $f_\theta(\frac{1}{2}x^{(i)} + \frac{1}{2}x^{(i+1)}) = m^{(i)}$. Because P is convex and the hyperplanes are disjoint in P, the midpoint $x_m^{(i)} = \frac{1}{2}(x^{(i)} + x^{(i+1)})$ lies in $\operatorname{relint}(R_i)$. Thus, the pullback $f_\theta^{-1}(H')$ contains a point on the boundary $(x^{(i)})$ and a point in the interior $(x_m^{(i)})$ of R_i. This implies that $f_\theta^{-1}(H')$ and H_i intersect transversely. Specifically, for each i, the pullback $f_\theta^{-1}(H')$ contains a point on the boundary of R_i ($x^{(i)} \in H_i \cap \operatorname{relint}(P)$) and a point in the interior ($x_m^{(i)} \in \operatorname{relint}(R_i)$). Since the pullback $f_\theta^{-1}(H')$ on R_i is an affine hyperplane and contains a point in $\operatorname{relint}(R_i)$, it cannot be identical to the boundary hyperplane H_i. Consequently, on R_i, we have that $f_\theta^{-1}(H')$ and H_i are distinct hyperplanes sharing at least one point in $\operatorname{relint}(P)$, ensuring they intersect transversely in P, that is, their intersection in P is of dimension $\dim(P) - 2$.

We next show that $f_\theta^{-1}(H') \cap P$ is connected. For each $i \in \{0, \dots, n\}$, the set $f_\theta^{-1}(H') \cap R_i$ is the intersection of the polyhedron R_i with an affine hyperplane, since f_θ is affine-linear on R_i. Hence it is convex, and therefore connected whenever nonempty. We have already shown that it is nonempty for every i. Moreover, for each $i \in [n]$, the set $f_\theta^{-1}(H')$ contains the point $x^{(i)} \in H_i \cap P$, so $f_\theta^{-1}(H') \cap R_{i-1}$ and $f_\theta^{-1}(H') \cap R_i$ meet along the common boundary hyperplane $H_i \cap P$. Thus the connected sets $f_\theta^{-1}(H') \cap R_0$, $f_\theta^{-1}(H') \cap R_1$, $\dots$, $f_\theta^{-1}(H') \cap R_n$ form a chain with consecutive intersections nonempty. Their union is therefore connected. Since this union equals $f_\theta^{-1}(H') \cap P$, the latter is connected.

Finally, since f_θ is linear on R_{n-1} (a polyhedron), it maps points in the relative interior of R_{n-1} to points in the relative interior of $f_\theta(R_{n-1})$. Since $x_m^{(n-1)} \in \operatorname{relint}(R_{n-1})$, we have $f_\theta(x_m^{(n-1)}) \in \operatorname{relint} f_\theta(R_{n-1})$. The latter is $m^{(n-1)} \in H'$ and thus H' intersects $\operatorname{relint}(f_\theta(R_{n-1}))$. Moreover, since $f_\theta^{-1}(H')$ and H_{n-1} intersect transversely in P, we have $f_\theta(R_{n-1}) \not\subseteq H'$. Thus $\dim(f_\theta(R_{n-1}) \cap H') = \dim(f_\theta(R_{n-1})) - 1$ and hence H' is inside $f_\theta(R_{n-1})$ in the sense of Definition 4.1.

□

Lemma 4.7. *Let $P \subseteq \mathbb{R}^d$ be a polyhedron, let $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^n$ be a ReLU layer, and let $H' \subseteq \mathbb{R}^n$ be a hyperplane such that $f_\theta^{-1}(H') \cap P$ is connected and, for all $H \in \mathcal{H}_\theta$, the hypersurfaces $f_\theta^{-1}(H')$ and H intersect transversely in P. Then there exists $\varepsilon > 0$ such that for every hyperplane H'' that is ε-close to H',*

1. *$f_\theta^{-1}(H'') \cap P$ is connected, and*
2. *$f_\theta^{-1}(H'')$ and H intersect transversely in P for all $H \in \mathcal{H}_\theta$.*

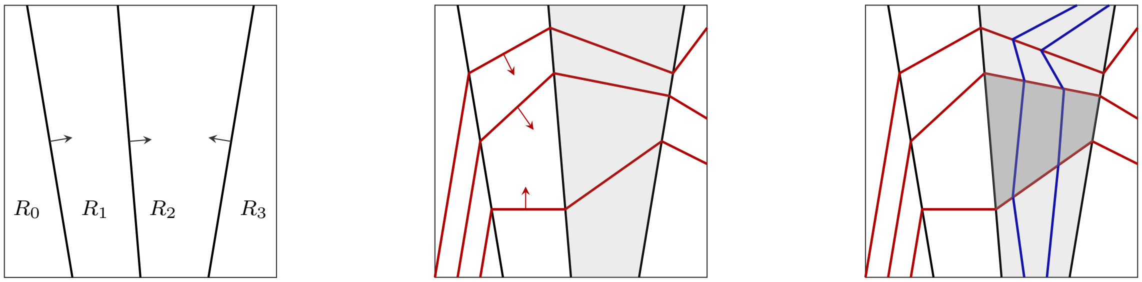
Proof. By the definition of transversality in P (Definition 3.26), for each $H \in \mathcal{H}_\theta$, there exists an intersection point $x \in f_\theta^{-1}(H') \cap H \cap \operatorname{relint}(P)$, where each local normal of $f_\theta^{-1}(H')$ and the normal of H are linearly independent. Because linear independence is an open condition and the intersection point x varies continuously with the parameters (w', b') corresponding to the hyperplane H', there exists a neighborhood in the parameter space such that these properties are preserved.

According to our definition of ε-closeness, this neighborhood corresponds to a ball B_ε around $(w', b')/\|w'\|$ or $(-w', -b')/\|w'\|$. Since $\mathcal{H}_\theta$ is finite, we can choose $\varepsilon > 0$ small enough to satisfy these conditions for all H simultaneously. For any H'' that is ε-close to H', the pullback $f_\theta^{-1}(H'')$ maintains a non-empty intersection with each $H \cap \operatorname{relint}(P)$ and preserves the linear independence of local normal vectors, ensuring the intersection remains transverse in P. By the same argument as in the proof of Lemma 4.6, $f_\theta^{-1}(H'') \cap P$ is connected. □

4.3 Pulling cTPIC Back to Input Space

Lemma 4.6 and Lemma 4.5 give us a recipe to iteratively construct slab layers $f_{\theta_1}, \dots, f_{\theta_L}$ such that $f_{\theta_{\ell+1}} \circ f_{\theta_\ell}$ satisfy cTPIC on a (possibly) lower-dimensional polytope $Q^{(\ell)}$ in $\mathbb{R}^{n_{\ell-1}}$. Next, we aim to pull the transverse intersections back to the input space, for which we need the following definition.

Definition 4.8. Let $T \colon \mathbb{R}^d \to \mathbb{R}^n$ an affine linear map given by $x \mapsto Ax + b$, and let $B \subseteq \mathbb{R}^n$ be a piecewise linear hypersurface. Then we say that T is *transverse* to B if for every local normal w of B we have that $A^\top w \neq 0$ and hence $T^{-1}(B)$ is a piecewise linear hypersurface.



(a) A generic oriented slab layer f_{θ_1} with 3 neurons on a polytope (the square). The layer's hyperplanes partition the polytope into 4 regions.

(b) The gray area is $P^{(1)}$. The red bent hyperplanes are the pullbacks of the hyperplanes of the slab layer f_{θ_2} on $Q^{(1)} = f_{\theta_1}(P^{(1)})$ along f_{θ_1}.

(c) The dark gray area is $P^{(2)}$. The blue bent hyperplanes are the pullbacks of the hyperplanes of the slab layer f_{θ_3} on $Q^{(2)}$ along $f_{\theta_2} \circ f_{\theta_1}$.

Figure 4: Illustration of the inductive construction in Theorem 4.10 that satisfies cTPIC and LRA.

With this definition at hand, we show that, generically, these intersections can be pulled back to the input space.

Lemma 4.9. *Let $P \subseteq \mathbb{R}^d$ be a full-dimensional polyhedron, let $T\colon \mathbb{R}^d \to \mathbb{R}^n$ be an affine map, and let $f_\theta\colon \mathbb{R}^n \to \mathbb{R}^m$ be an oriented slab layer on $T(P)$. Let $H' \subseteq \mathbb{R}^m$ be a hyperplane such that, for every $H \in \mathcal{H}_\theta$, $f_\theta^{-1}(H')$ and H intersect transversely in $T(P)$. Furthermore, assume that T is transverse to both $f_\theta^{-1}(H')$ and H. Then $(f_\theta \circ T)^{-1}(H')$ and $T^{-1}(H)$ intersect transversely in P.*

Proof. Let $H \in \mathcal{H}_\theta$ be a hyperplane in $\mathbb{R}^n$ and let $B = f_\theta^{-1}(H')$ be the piecewise linear hypersurface in $\mathbb{R}^n$. By the first assumption, there exists a point $y \in B \cap H \cap \operatorname{relint}(T(P))$. Since affine maps send relative interiors to relative interiors of their images, and $y \in \operatorname{relint}(T(P))$, there exists a point $x \in \operatorname{relint}(P)$ such that $T(x) = y$. By the definition of the pullback, $x \in (f_\theta \circ T)^{-1}(H')$ and $x \in T^{-1}(H)$, which establishes that the intersection in the relative interior of P is non-empty.

To show that the intersection is transverse in P, we analyze the local linear parts. Let w_H be the normal vector of H and w_B be any local normal vector of B at y. Let A be the linear part of the affine map T and $V = \operatorname{im}(A)$ be its image. The affine hull $\operatorname{aff}(T(P))$ is a translate of V. The assumption that H and B intersect transversely in $T(P)$ implies that their projections onto V, denoted $\pi_V(w_H)$ and $\pi_V(w_B)$, are linearly independent.

The local normal vectors of the pulled-back hypersurfaces at x are $u = A^\top w_H$ and $v = A^\top w_B$. Suppose there exist $\alpha, \beta \in \mathbb{R}$ such that $\alpha u + \beta v = 0$. This is equivalent to $A^\top(\alpha w_H + \beta w_B) = 0$. Let $z = \alpha w_H + \beta w_B$. The condition $A^\top z = 0$ implies that $z \in V^\perp$, or equivalently, $\pi_V(z) = 0$. By the linearity of the projection, we have:

$$\alpha \pi_V(w_H) + \beta \pi_V(w_B) = 0.$$

Since $\pi_V(w_H)$ and $\pi_V(w_B)$ are linearly independent in V, we must have $\alpha = \beta = 0$. Consequently, the pulled-back normals u and v are linearly independent in $\mathbb{R}^d$.

Since the intersection is non-empty in $\operatorname{relint}(P)$ and the local normal vectors are linearly independent, the intersection $(f_\theta \circ T)^{-1}(H') \cap T^{-1}(H) \cap P$ is pure of dimension $\dim(P) - 2$. This satisfies the definition of transversality in P. □

4.4 The Inductive Construction

Combining Lemma 4.5, Lemma 4.6 and Lemma 4.9, we establish in Theorem 4.10 the existence of an open set of parameters satisfying both cTPIC and LRA. Our construction proceeds inductively, layer by layer, as illustrated in Figure 4.

Theorem 4.10. *Let $\mathcal{A} = (n_0, \dots, n_{L+1})$ be any architecture with $n_i \geq 2$ for all $i \leq L$ and $P \subseteq \mathbb{R}^{n_0}$ a full-dimensional polytope. Then there is an open set $U \subseteq \Theta_{\mathcal{A}}$ such that f_θ satisfies cTPIC and LRA on P for all $\theta \in U$.*

Proof. If $L = 1$, there is no adjacent pair of hidden layers, so cTPIC is vacuous. Choose the first hidden layer to be a generic oriented slab layer on P, and choose the output weights generically so that every first-layer hyperplane carries nonzero tropical weight. Then LRA holds on P by Lemma 3.8. These conditions are open, so the desired open set exists.

Assume now that $L \geq 2$. We iteratively construct a sequence of full-dimensional polytopes $P = P^{(0)} \supseteq P^{(1)} \supseteq \cdots \supseteq P^{(L-2)}$ and a generic parameter θ satisfying LRA on P and such that the connected pieces of the bent hyperplanes from layers ℓ and $\ell + 1$ intersect transversely in $P^{(\ell-1)}$. The construction is illustrated in Figure 4.

Inductive construction We show the following by induction. For any $\ell \geq 2$, there are full-dimensional polytopes $P = P^{(0)} \supseteq \cdots \supseteq P^{(\ell-1)}$ and a generic parameter $\theta^{(\ell)} = (\theta_1, \dots, \theta_\ell) \in \Theta^{(\ell)}$ such that:

1. $f_{\theta^{(\ell-1)}}$ is affine linear on $P^{(\ell-1)}$
2. f_{θ_ℓ} is a generic oriented slab layer on $Q^{(\ell-1)} := f_{\theta^{(\ell-1)}}(P^{(\ell-1)})$ (Definition 4.1) and transparent on $X^{(\ell-1)} = f_{\theta^{(\ell-1)}}(P)$ (Definition 3.9)
3. $f_{\theta_\ell} \circ f_{\theta_{\ell-1}}$ satisfies cTPIC on $Q^{(\ell-2)}$ (Definition 3.26)
4. The affine linear map $T^{(\ell-2)} := f_{\theta^{(\ell-2)}}|_{P^{(\ell-2)}}$ (with $T^{(0)} := \mathrm{Id}|_P$ if $\ell = 2$) is transverse to the bent hyperplanes of $f_{\theta_\ell} \circ f_{\theta_{\ell-1}}$ (Definition 4.8)

Moreover, f_{θ_1} is a generic oriented slab layer on $Q^{(0)} := P^{(0)} = P$ transparent on $X^{(0)} := P^{(0)} = P$.

Before we start the induction, we note that the fourth point is true for generic parameters. The linear part of the map $T^{(\ell-2)} = f_{\theta^{(\ell-2)}}|_{P^{(\ell-2)}}$ is given by the matrix product $A = D_{S_{\ell-2}} W^{(\ell-2)} D_{S_{\ell-3}} \cdots D_{S_1} W^{(1)}$. The (local) normals of the (bent) hyperplanes of $f_{\theta_\ell} \circ f_{\theta_{\ell-1}}$ are given by the vectors $v_j = W_j^{(\ell-1)}$ and vectors $u_{S,i} = (W^{(\ell)} D_S W^{(\ell-1)})_i$ for $j \in [n_{\ell-1}], i \in [n_\ell]$ and $S \subseteq [n_{\ell-1}]$. By definition, $T^{(\ell-2)}$ is transverse to these bent hyperplanes if $v_j A \neq 0$ and $u_{S,i} A \neq 0$ for all possible choices $j \in [n_{\ell-1}], i \in [n_\ell], S \subseteq [n_{\ell-1}]$. If $v_j A = 0$, then $D_{\{j\}} W^{(\ell-1)} D_{S_{\ell-2}} W^{(\ell-2)} \cdots W^{(1)} = 0$, which would contradict Condition 2 of Definition 3.1, that states that the matrix product $D_{\{j\}} W^{(\ell-1)} D_{S_{\ell-2}} W^{(\ell-2)} \cdots W^{(1)}$ must achieve the maximum possible rank. In the same manner, if $u_{S,i} A = 0$, then $D_{\{i\}} W^{(\ell)} D_{S_{\ell-1}} W^{(\ell-1)} \cdots W^{(1)} = 0$, which again would contradict Definition 3.1. Consequently, for any generic parameter, $T^{(\ell-2)}$ is necessarily transverse to the bent hyperplanes of the subsequent layer.

Base case of the induction For the base case of the induction, let $\ell = 2$. The full-dimensional polytope P has dimension $n_0 \geq 2$, and we can pick a hyperplane H inside P (Definition 4.1). Now by Lemma 4.5, there is an oriented generic slab layer $f_{\theta_1} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_1}$ on P that is transparent on P. By Lemma 4.6, there is a hyperplane $H' \subseteq \mathbb{R}^{n_1}$ such that $f_{\theta_1}^{-1}(H')$ intersects every $H \in \mathcal{H}_{\theta_1}$ transversely and $f_{\theta_1}^{-1}(H') \cap P$ is connected. By Lemma 4.7, there is a $\varepsilon_1 > 0$ such that for every hyperplane H'' that is ε_1-close to H', the bent hyperplane $f_{\theta_1}^{-1}(H'')$ still has transverse intersection with each $H \in \mathcal{H}_{\theta_1}$ and is still connected in P. Moreover, again by Lemma 4.6, there is a region $P^{(1)} \in \mathcal{C}_{\theta_1}(P)$ such that H' is inside $Q^{(1)} = f_{\theta_1}(P^{(1)})$. By Remark 4.4, there is an ε_2 such that every hyperplane H'' that is ε_2-close to H' is also inside $Q^{(1)}$. Let $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$. Now, by Lemma 4.5, there is a generic oriented slab layer f_{θ_2} on $Q^{(1)}$ that is transparent on the bounded set $X^{(1)} = f_{\theta_1}(P)$ and ε-close to H'. Finally, a small perturbation of the construction ensures that the parameter is generic, implying that the fourth point holds. This completes the base case.

Induction step Now for the induction step, let $\ell \geq 2$ and assume that the induction hypothesis holds. We proceed to construct layer $\ell + 1$ analogously to the base case.

By the induction hypothesis, f_{θ_ℓ} is a generic oriented slab layer on $Q^{(\ell-1)}$. Thus, by Lemma 4.6, there exists a hyperplane $H' \subseteq \mathbb{R}^{n_\ell}$ and a region $R \in \mathcal{C}_{\theta_\ell}(Q^{(\ell-1)})$ such that: $f_{\theta_\ell}^{-1}(H')$ is connected in $Q^{(\ell-1)}$, intersects every $H \in \mathcal{H}_{\theta_\ell}$ transversely in $Q^{(\ell-1)}$, and such that H' is inside $f_{\theta_\ell}(R)$.

We define $P^{(\ell)} := f_{\theta^{(\ell-1)}}^{-1}(R)$, which is a region in $\mathcal{C}_{\theta^{(\ell)}}(P^{(\ell-1)})$ and hence has dimension $\dim(P^{(\ell-1)}) = \dim(P)$. We have $Q^{(\ell)} := f_{\theta^{(\ell)}}(P^{(\ell)}) = f_{\theta_\ell}(R)$. By Lemma 4.5, there exists a generic oriented slab layer $f_{\theta_{\ell+1}}$ on $Q^{(\ell)}$ that is transparent on the bounded set $X^{(\ell)} := f_{\theta_\ell}(X^{(\ell-1)}) = f_{\theta^{(\ell)}}(P)$ and whose hyperplane arrangement $\mathcal{H}_{\theta_{\ell+1}}$ is ε-close to H', for any choice of $\varepsilon > 0$. We set $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$, where $\varepsilon_1 > 0$ is given by Lemma 4.7 and ensures that, for every H'' that is ε_1-close to H', $f_{\theta_\ell}^{-1}(H'')$ is still connected in $Q^{(\ell-1)}$, intersects every $H \in \mathcal{H}_{\theta_\ell}$ transversely in $Q^{(\ell-1)}$, and $\varepsilon_2 > 0$ is given by Remark 4.4 and ensures that every H'' that is ε_2-close to H' is inside $Q^{(\ell)}$. By this construction, we have:

- $f_{\theta^{(\ell)}}$ is affine-linear on $P^{(\ell)}$;
- $f_{\theta_{\ell+1}}$ satisfies the generic oriented slab and transparency requirements on $Q^{(\ell)}$ and $X^{(\ell)}$, respectively;
- The ε-closeness of $\mathcal{H}_{\theta_{\ell+1}}$ to H' ensures that $f_{\theta_{\ell+1}} \circ f_{\theta_\ell}$ satisfies cTPIC on $Q^{(\ell-1)}$;
- If necessary, we can slightly perturb the construction to ensure that the resulting parameter is generic, which guarantees that $T^{(\ell-1)} = f_{\theta^{(\ell-1)}}|_{P^{(\ell-1)}}$ is transverse to the bent hyperplanes of $f_{\theta_{\ell+1}} \circ f_{\theta_\ell}$.

This concludes the induction step.

The constructed parameters satisfy cTPIC and LRA We first verify cTPIC on the input polytope P. Fix a hidden neuron (i, ℓ) with $\ell \in [L]$. If $\ell = 1$, we set $C_{i,1} := B_{i,1}(\theta) \cap P$. Since f_{θ_1} is a slab layer on P, this is the intersection of P with an affine hyperplane, hence a connected piecewise linear hypersurface.

Now let $\ell \geq 2$. By Property 3, the two-layer network $f_{\theta_\ell} \circ f_{\theta_{\ell-1}}$ satisfies cTPIC on $Q^{(\ell-2)}$. Hence, for the neuron (i, ℓ), there exists a connected piecewise linear hypersurface $\widetilde{C}_{i,\ell} \subseteq B_{i,\ell}(\theta) \cap Q^{(\ell-2)}$. We define $C_{i,\ell} := \big(f_{\theta^{(\ell-2)}}|_{P^{(\ell-2)}}\big)^{-1}(\widetilde{C}_{i,\ell}) \subseteq B_{i,\ell}(\theta) \cap P$. By Property 1, the map $f_{\theta^{(\ell-2)}}|_{P^{(\ell-2)}}$ is affine-linear. Since $P^{(\ell-2)}$ is convex, the preimage of a connected subset of its image under this affine map is connected. Hence $C_{i,\ell}$ is connected.

Now let (j, ℓ) and $(i, \ell+1)$ be neurons in adjacent layers. By Property 3, the two-layer network $f_{\theta_{\ell+1}} \circ f_{\theta_\ell}$ satisfies cTPIC on $Q^{(\ell-1)}$, so the corresponding connected hypersurface pieces intersect transversely there. By Property 4 and Lemma 4.9, these intersections pull back along $f_{\theta^{(\ell-1)}}|_{P^{(\ell-1)}}$ to transverse intersections between $C_{j,\ell}$ and $C_{i,\ell+1}$ in $P^{(\ell-1)} \subseteq P$. Thus the family $\{C_{i,\ell}\}_{i,\ell}$ witnesses that f_θ satisfies cTPIC on P.

Moreover, Property 2 shows that f_{θ_ℓ} is transparent on $X^{(\ell-1)} = a^{(\ell-1,\theta)}(P)$ for every $\ell \in [L]$. Since θ is generic, Lemma 3.10 implies that θ satisfies LRA on P.

Finally, the construction is stable under sufficiently small perturbations: transparency, the slab-layer conditions, connectedness of the chosen pullbacks, and the required transverse intersections are all open conditions. Hence we obtain an open subset of parameters satisfying cTPIC and LRA on P. This concludes the proof. ☐

4.5 Functional Dimension and Identifiability

By Theorem 3.31, generic parameters satisfying cTPIC and LRA are identifiable among generic parameters. On the resulting open set of generic parameters, the quotient realization map attains maximal rank, and therefore the dimension of the function space equals the quotient parameter dimension.

Theorem 4.11. *Let $\mathcal{A} = (n_0, \ldots, n_{L+1})$ with $n_i \geq 2$ for all $i \leq L$. Then there exists a nonempty open set $U_0 \subseteq \widetilde{\Theta}_{\mathcal{A}}$ such that every $\theta \in U_0$ is identifiable among generic parameters. Consequently,* $\dim(\mu_{\mathcal{A}}(\Theta_{\mathcal{A}})) = \sum_{\ell=1}^{L+1} n_\ell n_{\ell-1} + n_{L+1}$.

Proof. By Theorem 4.10, there exists a nonempty open set $U \subseteq \Theta_{\mathcal{A}}$ such that every $\theta \in U$ satisfies cTPIC and LRA on a full-dimensional polytope. Intersecting with the open dense generic locus $\widetilde{\Theta}_{\mathcal{A}}$, we obtain a nonempty open set $U_0 \subseteq \widetilde{\Theta}_{\mathcal{A}}$ such that every $\theta \in U_0$ is identifiable among generic parameters by Theorem 3.31.

Modulo permutation and positive scaling symmetries, the realization map restricted to U_0 therefore has a trivial fiber on U_0. Hence the induced quotient realization map attains rank $\dim(\overline{\Theta}_{\mathcal{A}}) = \sum_{\ell=1}^{L+1}(n_\ell n_{\ell-1} + n_\ell) - \sum_{\ell=1}^{L} n_\ell$ (see, e.g., Grigsby et al., 2023). This shows that $\dim(\mu_{\mathcal{A}}(\Theta_{\mathcal{A}})) \geq \dim(\overline{\Theta}_{\mathcal{A}})$. The reverse inequality is automatic, since the image of the realization map cannot have larger dimension than the quotient parameter space. Therefore equality holds, and expanding the right-hand side gives $\dim(\mu_{\mathcal{A}}(\Theta_{\mathcal{A}})) = \sum_{\ell=1}^{L+1} n_\ell n_{\ell-1} + n_{L+1}$. □

The preceding theorem still yields identifiability only among the generic parameters. To upgrade this to unconditional identifiability, we exclude functions that admit nongeneric realizations. Since the nongeneric locus is a proper algebraic subset of parameter space, its image in function space is lower-dimensional than the full image established above.

Lemma 4.12. *Let $N = \Theta_{\mathcal{A}} \setminus \widetilde{\Theta}_{\mathcal{A}}$ and let $q : \Theta_{\mathcal{A}} \to \overline{\Theta}_{\mathcal{A}}$ be the quotient map. Then $\dim q(N) < \dim \overline{\Theta}_{\mathcal{A}}$.*

Proof. The generic locus $\widetilde{\Theta}_{\mathcal{A}}$ is invariant under permutation and positive scaling symmetries, hence so is N. Since $\widetilde{\Theta}_{\mathcal{A}}$ is open dense in $\Theta_{\mathcal{A}}$ and q is an open map, $q(\widetilde{\Theta}_{\mathcal{A}})$ is open dense in $\overline{\Theta}_{\mathcal{A}}$. Therefore $q(N) = \overline{\Theta}_{\mathcal{A}} \setminus q(\widetilde{\Theta}_{\mathcal{A}})$ has empty interior. Since $q(N)$ is semialgebraic, it has dimension strictly smaller than $\overline{\Theta}_{\mathcal{A}}$. □

Theorem 4.13. *Let $\mathcal{A} = (n_0, \ldots, n_{L+1})$ with $n_i \geq 2$ for all $i \leq L$. Then there exists a nonempty open set $U \subseteq \Theta_{\mathcal{A}}$ such that every $\theta \in U$ is identifiable.*

Proof. Let $U_0 \subseteq \widetilde{\Theta}_{\mathcal{A}}$ be the open set from Theorem 4.11, so that every $\theta \in U_0$ is identifiable among generic parameters. and let $N := \Theta_{\mathcal{A}} \setminus \widetilde{\Theta}_{\mathcal{A}}$ the nongeneric locus.

Let $q : \Theta_{\mathcal{A}} \to \overline{\Theta}_{\mathcal{A}}$ be the quotient map. By Lemma 4.12, $\dim q(N) < \dim \overline{\Theta}_{\mathcal{A}}$. Since $\mu_{\mathcal{A}} = \overline{\mu}_{\mathcal{A}} \circ q$, we get $\dim \mu_{\mathcal{A}}(N) \leq \dim q(N) < \dim \overline{\Theta}_{\mathcal{A}}$. By Theorem 4.11, $\dim(\mu_{\mathcal{A}}(\Theta_{\mathcal{A}})) = \dim(\overline{\Theta}_{\mathcal{A}})$, so $\mu_{\mathcal{A}}(N)$ is a proper lower-dimensional subset of the full function-space image.

Now consider $B := U_0 \cap \mu_{\mathcal{A}}^{-1}(\mu_{\mathcal{A}}(N))$. Since $\mu_{\mathcal{A}}(N)$ is lower-dimensional whereas $\mu_{\mathcal{A}}(U_0)$ has full dimension, the set B cannot contain a nonempty open subset of U_0.Since B has empty interior in U_0, choose a nonempty open subset $U \subseteq U_0 \setminus B$.

We claim that every $\theta \in U$ is identifiable. Let $\theta \in U$, and let $\eta \in \Theta_{\mathcal{A}}$ satisfy $f_\eta = f_\theta$. If η were nongeneric, then $f_\theta = f_\eta \in \mu_{\mathcal{A}}(N)$, which would imply $\theta \in B$, contradicting $\theta \in U$. Thus η must be generic. Since $\theta \in U_0$ and every point of U_0 is identifiable among generic parameters, it follows that $\eta \sim \theta$. Therefore θ is identifiable. □

Remark 4.14. Theorem 4.13 leaves open the existence of identifiable parameters in architectures that contain non-output layers of width 1. If a layer ℓ consists of a single neuron, then the activation image $a^{(\ell,\theta)}(\mathbb{R}^d)$ is at most one-dimensional. Consequently, the Transverse Pairwise Intersection Condition (TPIC) cannot be satisfied for any neurons in subsequent layers $k > \ell$, as their preactivations depend on a single direction. While the locations of breakpoints along this single direction are informative about the weights, the absence of transverse intersections appears to prevent the unique assignment of neurons to breakpoints (modulo within-layer permutations). By adapting the construction in Lemma 4.5 to this setting, it appears possible to construct parameters for which the breakpoints are fixed while their neurons of origin remain ambiguous. This motivates the following conjecture.

Conjecture 1.

1. *There exist architectures $\mathcal{A} = (n_0, \ldots, n_{L+1})$ with $n_\ell = 1$ for some $\ell \leq L$ for which no identifiable parameters exist.*

2. *For every architecture, there exists a parameter that is* finitely identifiable, *meaning that its fiber in the quotient space $\overline{\Theta_{\mathcal{A}}}$ is a finite set.*

Conjecture 1 1) is motivated by the observation that TPIC fails for architectures containing layers of width 1; however, it remains somewhat speculative, and we hope that its resolution will stimulate the development of new methods for handling such cases.

If Conjecture 1 2) were to hold on an open subset of parameters, it would follow that architectures containing layers of width 1 also attain the expected dimension, namely the number of parameters minus the number of hidden neurons. Indeed, a finite fiber implies that the realization map is locally an embedding.

5 Minimality Does Not Imply Identifiability

Note that in a ReLU network identifiability implies minimality. Indeed, identifiability implies that there is no other parameter zeroing out neurons that represents the same function.

In this subsection, we use the dependency graph and a modification of our construction from Theorem 4.10 to show that minimal generic parameters do not need to be identifiable. Moreover, even if a network has an open subset of identifiable parameters it can still have an open subset of minimal parameters that are not identifiable. This stands in contrast to polynomial networks, where the identifiability of a generic parameter implies the identifiability of all generic parameters (see, e.g., Usevich et al., 2025, Finkel et al., 2025, Shahverdi et al., 2026b). More broadly, this example highlights that, in the ReLU setting, the relation between (non-)identifiability and reducibility of the architecture is subtle, a theme that also arises in recent abstract work connecting identifiability and equivariance in deep networks (Shahverdi et al., 2026a).

We first rigidify the prefix of the network using the dependency graph, then attach an additional layer with a two-neuron always-active linear block producing a 2-dimensional fiber, and finally use a dimension comparison with strict subarchitectures to deduce minimality on an open set.

5.1 Embedding an Identifiable Prefix

The following lemma states a condition under which we can embed an identifiable parameter into a deeper network such that the embedded part of the parameter remains identifiable in the deeper network.

Lemma 5.1. *Let $\mathcal{A} = (n_0, \ldots, n_L, m)$ be an architecture, and let $\theta \in \tilde{\Theta}_{\mathcal{A}}$. Decompose the network function as $f_\theta = f_{\theta_L,\theta_{L+1}} \circ f_{\theta^{(L-1)}}$, where $f_{\theta^{(L-1)}} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L-1}}$ and $f_{\theta_L,\theta_{L+1}} : \mathbb{R}^{n_{L-1}} \to \mathbb{R}^m$. Suppose that θ satisfies:*

1. *$\theta^{(L-1)}$ satisfies cTPIC and LRA on some polytope $P \subseteq \mathbb{R}^{n_0}$,*
2. *f_{θ_L} is transparent on $f_{\theta^{(L-1)}}(P)$, and*
3. *at least one bent hyperplane $B_{L,j}(\theta)$ intersects transversely with all hyperplanes $B_{L-1,i}(\theta)$ in P.*

Then, for each generic $\eta = (\eta^{(L-1)}, \eta_L, \eta_{L+1})$ with $f_\eta = f_\theta$, the truncated parameters $\eta^{(L-1)}$ and $\theta^{(L-1)}$ are equivalent up to the trivial symmetries of the sub-architecture $(n_0, \ldots, n_{L-1})$.

Proof. Let G' be the dependency graph arising from the first $L-1$ layers of θ on P. Since $\theta^{(L-1)}$ satisfies cTPIC and LRA on P, Lemma 3.29 implies that G' contains a layered subgraph with $L-1$ levels corresponding to the hidden layers $1, \ldots, L-1$, and with all adjacent levels fully connected.

By assumption (2), the layer f_{θ_L} is transparent on $f_{\theta^{(L-1)}}(P)$, so the visible breakpoint geometry created by the first $L-1$ layers is preserved in the full function f_θ. Hence the same layered subgraph G' is contained in the dependency graph of f_θ.

By assumption (3), there exists a bent hyperplane $B_{L,j}(\theta)$ in the last hidden layer intersecting transversely every bent hyperplane $B_{L-1,i}(\theta)$ in P. Thus the dependency graph of f_θ contains a vertex v_L corresponding to $B_{L,j}(\theta)$ together with edges from every vertex in the $(L-1)$-st level of G' to v_L.

Now let η be generic with $f_\eta = f_\theta$. By Proposition 3.25, there is an order-preserving map from the candidate bent hyperplanes of f_θ to the hidden neurons of η. Applied to the layered subgraph G', this forces its $L-1$ levels to occupy $L-1$ distinct hidden layers of η. Since v_L lies strictly above the last level of G', it must occupy a strictly later hidden layer. As η has only L hidden layers, the vertices of G' must therefore occupy exactly the first $L-1$ hidden layers of η.

Hence the visible breakpoint geometry generated by the prefix $\theta^{(L-1)}$ is realized in the first $L-1$ layers of η. Since $\theta^{(L-1)}$ satisfies cTPIC and LRA on P, the reverse-engineering theorem Theorem 3.28 recovers these first $L-1$ layers up to permutation and positive scaling. As in the proof of Theorem 3.31, generically, there is no sign ambiguity.

□

Note that in contrast to cTPIC, the condition in Lemma 5.1 does not require that every bent hyperplane from layer $L-1$ to be intersected by every bent hyperplane from layer L. Consequently, Theorem 3.31 does not apply to the parameters of the last hidden layer, which therefore need not be identifiable.

5.2 Add a Linear Block

The strategy to construct a minimal but non-identifiable parameter is to compose an identifiable parameter (as in Theorem 4.10) with a one-hidden-layer network that is minimal but has a continuous parameter symmetry, since some of its neurons do not introduce any breakpoints on the image of the preceding layer.

Proposition 5.2. *For architectures $\mathcal{A} = (n_0, \dots, n_L, n_{L+1})$ with $L \geq 2$ and $n_i \geq 2$ for $i \in \{0, \dots, L+1\}$ and $n_L \geq 4$, there exists a nonempty open subset $U \subseteq \Theta_{\mathcal{A}}$ such that for every $\theta \in U$, $\dim\big(\overline{\mu}_{\mathcal{A}}^{-1}(f_\theta)\big) \geq 2$, and $\dim \mu_{\mathcal{A}}(U) = \dim \mu_{\mathcal{A}}(\Theta_{\mathcal{A}}) - 2$.*

Proof. Let $P \subseteq \mathbb{R}^{n_0}$ be a full-dimensional polytope. We construct a parameter $\theta \in \Theta_{\mathcal{A}}$ in two stages.

By Theorem 4.10, there exists a generic parameter $\theta^{(L-1)} \in \Theta^{(L-1)}$ for the architecture $(n_0, \dots, n_{L-1})$ such that $f_{\theta^{(L-1)}}$ satisfies cTPIC and LRA on P. In particular, $\theta^{(L-1)}$ is identifiable among generic parameters. Let $Q^{(L-1)} := f_{\theta^{(L-1)}}(P) \subseteq \mathbb{R}^{n_{L-1}}$.

We also note that, by the slab-layer construction, the prefix image affinely spans the last latent space: $\operatorname{aff}\big(f_{\theta^{(L-1)}}(\mathbb{R}^{n_0})\big) = \mathbb{R}^{n_{L-1}}$. Indeed, the last layer of the prefix is an oriented slab layer. As in the proof of Lemma 4.6, the images of points on its n_{L-1} slab hyperplanes are linearly independent in $\mathbb{R}^{n_{L-1}}$, and the image of one slab region contains a point outside their affine span. Hence the prefix image contains $n_{L-1}+1$ affinely independent points. Set $Y := f_{\theta^{(L-1)}}(\mathbb{R}^{n_0})$.

For the last hidden layer and the output layer, partition the neurons of layer L into $J := [n_L - 2]$, and $K := \{n_L - 1, n_L\}$. We think of the neurons in J as *bending neurons* and those in K as *linear neurons*.

Choose the parameters $(W_J^{(L)}, b_J^{(L)})$ of the bending neurons so that they form a generic oriented slab layer on $Q^{(L-1)}$ intersecting all the bent hyperplanes from layer $L-1$. Since $n_L - 2 \geq 2$, this can be achieved by the same slab-layer construction used in Theorem 4.10.

Next choose the parameters $(W_K^{(L)}, b_K^{(L)})$ of the two linear neurons so that their augmented input matrix $W_{\text{lin}} := \begin{bmatrix} W_K^{(L)} & b_K^{(L)} \end{bmatrix} \in \mathbb{R}^{2\times(n_{L-1}+1)}$ has strictly positive entries and rank 2. Since $Y \subseteq \mathbb{R}^{n_{L-1}}_{\geq 0}$, these two preactivations are strictly positive on Y. Hence the two neurons in K are active on the entire prefix image and contribute only an affine-linear map on the realized domain of the prefix. Choose also $V_{\text{lin}} := W_{:,K}^{(L+1)} \in \mathbb{R}^{n_{L+1}\times 2}$ generically of rank 2.

This completes the construction of θ.

Positive-dimensional fiber. The contribution of the two linear neurons is the affine map represented by $B_{\text{lin}} = V_{\text{lin}} W_{\text{lin}}$. For any $M \in \mathrm{GL}_2(\mathbb{R})$ sufficiently close to the identity such that MW_{lin} still has strictly positive entries, define $W'_{\text{lin}} = MW_{\text{lin}}$, and $V'_{\text{lin}} = V_{\text{lin}} M^{-1}$. Then the transformed preactivations are again strictly positive on Y, and the affine contribution of the two-neuron block is unchanged: $V'_{\text{lin}} W'_{\text{lin}} = V_{\text{lin}} M^{-1} M W_{\text{lin}} = V_{\text{lin}} W_{\text{lin}}$. Therefore this transformation produces a local $\mathrm{GL}_2(\mathbb{R})$-family of parameters realizing the same function. Modulo the usual positive scaling symmetry of the two neurons, this family has dimension $\dim(\mathrm{GL}_2(\mathbb{R})) - \dim((\mathbb{R}_{>0})^2) = 4 - 2 = 2$. Hence the fiber of f_θ in the quotient parameter space is positive-dimensional.

Image dimension. The properties used above are open:

- cTPIC and LRA for the prefix on P,

- transversality of the intersections between layers $L-1$ and L,
- affine spanning of the prefix image,
- strict positivity of the two linear neurons on the prefix image,
- the rank-2 conditions on W_{lin} and V_{lin}.

Hence these conditions persist on an open neighborhood U_0 of the constructed parameter θ, and for every $\eta \in U_0$ we still have $\dim\big(\overline{\mu}_{\mathcal{A}}^{-1}(f_\eta)\big) \geq 2$. Now let $U = U_0 \cap \tilde{\Theta}_{\mathcal{A}}$. We show that $\dim(\mu_{\mathcal{A}}(U)) = \dim(\mu_{\mathcal{A}}(\Theta_{\mathcal{A}})) - 2$.

Let $\eta, \eta' \in U$ satisfy $f_\eta = f_{\eta'}$. Since η, η' are generic and in U_0, by Lemma 5.1, the first $L-1$ layers of η and η' coincide up to the trivial symmetries. Likewise, by the reverse-engineering algorithm of Rolnick and Kording (2020) (Theorem 3.28), the parameters of the bending neurons in J coincide up to the trivial symmetries and finitely many sign choices. In particular, they contribute no positive-dimensional fiber directions. After shrinking U if necessary, the finite sign choices for the bending neurons are fixed, and the above local description applies uniformly on U.

It remains to control the always-active block. After applying the trivial symmetries to identify the first $L-1$ layers of η and η', let $Y_\eta := f_{\eta^{(L-1)}}(\mathbb{R}^{n_0})$. By the openness assumptions defining U_0, this set affinely spans $\mathbb{R}^{n_{L-1}}$. Therefore equality of realized functions forces the affine map contributed by the two always-active neurons to be fixed as an affine map on $\mathbb{R}^{n_{L-1}}$. Equivalently, in the notation above, any local variation in the fiber must satisfy $V_{\mathrm{lin}}^{\eta'} W_{\mathrm{lin}}^{\eta'} = V_{\mathrm{lin}}^{\eta} W_{\mathrm{lin}}^{\eta}$. Since both W_{lin}^{η} and V_{lin}^{η} have rank 2, all nearby rank-2 factorizations of this fixed matrix through a two-dimensional hidden space are of the form $W_{\mathrm{lin}}^{\eta'} = M W_{\mathrm{lin}}^{\eta}$, and $V_{\mathrm{lin}}^{\eta'} = V_{\mathrm{lin}}^{\eta} M^{-1}$ for some $M \in \mathrm{GL}_2(\mathbb{R})$. Thus, modulo the trivial symmetries, the only local continuous fiber directions in U are the two directions coming from the GL_2-factorization freedom of the always-active two-neuron block.

Hence the quotient realization map has local fiber dimension 2 on U. By the fiber-dimension theorem for semialgebraic maps, $\dim \mu_{\mathcal{A}}(U) = \dim(\overline{\Theta}_{\mathcal{A}}) - 2$. Since $\dim(\overline{\Theta}_{\mathcal{A}}) = \dim \mu_{\mathcal{A}}(\Theta_{\mathcal{A}})$, we conclude that $\dim \mu_{\mathcal{A}}(U) = \dim \mu_{\mathcal{A}}(\Theta_{\mathcal{A}}) - 2$. This completes the proof. ☐

The previous proposition isolates the two ingredients needed for the minimality argument. First, it produces an open set with positive-dimensional fibers, arising from the GL_2-symmetry of the always-active linear block. Second, it shows that the image of this family has codimension 2 inside the full function space of the ambient architecture.

5.3 Dimension Drop and Minimality

To obtain minimality, we now compare this image with the function spaces of strict subarchitectures. Since every strict subarchitecture has strictly smaller functional dimension, indeed, at least three dimensions less, the union of their images is lower-dimensional than the image of the constructed set. Removing this exceptional union therefore leaves a nonempty open subset of functions that are not realizable by any strict subarchitecture, while their fibers remain positive-dimensional.

Lemma 5.3. *Let $\mathcal{A} = (n_0, \ldots, n_L, n_{L+1})$ with $L \geq 2$ and $n_i \geq 2$ for all $i \leq L$. If $\mathcal{A}' = (n_0, n_1', \ldots, n_L', n_{L+1})$ is a strict subarchitecture of $\mathcal{A}$, then $\dim \mu_{\mathcal{A}'}(\Theta_{\mathcal{A}'}) \leq \dim \mu_{\mathcal{A}}(\Theta_{\mathcal{A}}) - 3$.*

Proof. It suffices to consider the case where $\mathcal{A}'$ is obtained from $\mathcal{A}$ by decreasing one hidden layer width n_k by one, since the general case follows by iterating this estimate.

So let $\mathcal{A}' = (n_0, \ldots, n_{k-1}, n_k - 1, n_{k+1}, \ldots, n_L, n_{L+1})$, for some $k \in [L]$. Then

$$\begin{aligned}\dim \mu_{\mathcal{A}}(\Theta_{\mathcal{A}}) - \dim \mu_{\mathcal{A}'}(\Theta_{\mathcal{A}'}) &\geq \big(n_k n_{k-1} + n_{k+1} n_k\big) - \big((n_k - 1)n_{k-1} + n_{k+1}(n_k - 1)\big) \\ &= n_{k-1} + n_{k+1}.\end{aligned}$$

The inequality stems from the fact that in case in $\mathcal{A}'$ we have a layer of width 1, then the expected functional dimension is only an upper bound. Since $n_{k-1} \geq 2$ and $n_{k+1} \geq 1$, we obtain $n_{k-1} + n_{k+1} \geq 3$, which proves the claim. ☐

Theorem 5.4. *For architectures $\mathcal{A} = (n_0, \ldots, n_L, n_{L+1})$ with $L \geq 2$ and $n_i \geq 2$ for $i \in \{0, \ldots, L+1\}$ and $n_L \geq 4$, there is a nonempty open subset $U \subseteq \Theta_\mathcal{A}$ such that every $\theta \in U$ is minimal but has a positive-dimensional fiber,* $\dim(\overline{\mu}_\mathcal{A}^{-1}(f_\theta)) > 0$.

Proof. By Proposition 5.2, there exists a nonempty open subset $U_0 \subseteq \Theta_\mathcal{A}$ such that for every $\theta \in U_0$, $\dim\big(\overline{\mu}_\mathcal{A}^{-1}(f_\theta)\big) \geq 2$, and $\dim \mu_\mathcal{A}(U_0) = \dim \mu_\mathcal{A}(\Theta_\mathcal{A}) - 2$. Let $\mathcal{R} := \bigcup_{\mathcal{A}' < \mathcal{A}} \mu_{\mathcal{A}'}(\Theta_{\mathcal{A}'}) \subseteq \mu_\mathcal{A}(\Theta_\mathcal{A})$, where the union runs over all strict subarchitectures $\mathcal{A}'$ of $\mathcal{A}$. By Lemma 5.3, each strict subarchitecture satisfies $\dim \mu_{\mathcal{A}'}(\Theta_{\mathcal{A}'}) \leq \dim \mu_\mathcal{A}(\Theta_\mathcal{A}) - 3$. Since there are only finitely many strict subarchitectures of $\mathcal{A}$, it follows that $\dim \mathcal{R} \leq \dim \mu_\mathcal{A}(\Theta_\mathcal{A}) - 3$. On the other hand, $\dim \mu_\mathcal{A}(U_0) = \dim \mu_\mathcal{A}(\Theta_\mathcal{A}) - 2$. Hence $\mathcal{R} \cap \mu_\mathcal{A}(U_0)$ is a lower-dimensional subset of $\mu_\mathcal{A}(U_0)$. Therefore there exists a nonempty relatively open subset $V \subseteq \mu_\mathcal{A}(U_0) \setminus \mathcal{R}$. Since V is relatively open in $\mu_\mathcal{A}(U_0)$, the preimage $U := U_0 \cap \mu_\mathcal{A}^{-1}(V)$ is a nonempty open subset of U_0. Now let $\theta \in U$. Since $f_\theta \in V \subseteq \mu_\mathcal{A}(U_0) \setminus \mathcal{R}$, the function f_θ is not realized by any strict subarchitecture of $\mathcal{A}$. Hence θ is minimal.

Moreover, because $U \subseteq U_0$, we still have $\dim\big(\overline{\mu}_\mathcal{A}^{-1}(f_\theta)\big) \geq 2 > 0$. Thus every parameter in U is minimal and has positive-dimensional fiber. ☐

While the proof above establishes minimality by a dimension argument and therefore does not require a more explicit structural analysis, Appendix B provides a complementary one-hidden-layer perspective explaining why the linear neurons in the last hidden layer are nonetheless necessary, despite not producing additional visible breakpoints.

6 Generic Depth Hierarchy

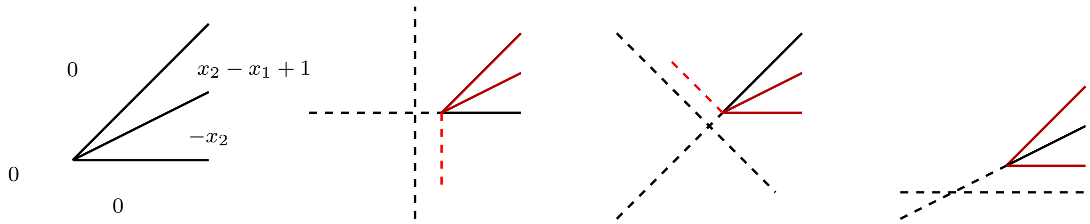
A significant challenge in neural network theory is to determine which functions are representable at fixed depth and whether there are functions representable by deep networks that are not at all representable by shallower ones, regardless of the width (Bakaev et al., 2026, Grillo et al., 2025, Averkov et al., 2025, Haase et al., 2023, Safran, 2026, Hertrich et al., 2023).

Most existing results and approaches rely on a representation of CPWL functions that reduces the depth separation question to determining the depth required to compute the maximum of d numbers. Our polyhedral framework, in contrast, allows us to study depth separation in the generic case. By leveraging the dependency graph as a functional invariant, we can show that, on an open subset of parameters, the layer hierarchy is structurally rigid. Unlike degenerate cases, where layer hierarchy might "collapse" due to exact cancellations or alignment of breakpoints, the functional geometry of a generic network encodes its depth. The following proposition formalizes this observation by showing that, for each architecture, there is an open set of parameters whose realized functions cannot be represented by any shallower architecture with generic parameters. For example, the function illustrated in Figure 4c cannot be represented with 2 hidden layers and generic weights. While a related statement is likely already implicit in the framework of Phuong and Lampert (2020), we find it useful to state it explicitly in this generality.

Proposition 6.1. *For any architecture $\mathcal{A} = (n_0, \ldots, n_L, m)$ with $n_i \geq 2$ for $i \in \{0, \ldots, L\}$, there exists a nonempty open subset of generic parameters U such that, for all $\theta \in U$, the function f_θ cannot be represented by any network of architecture $\mathcal{A}' = (n_0, n'_1, \ldots, n'_{L-1}, m)$ with generic parameters.*

Proof. Let θ be a parameter constructed in the proof of Theorem 4.10. By construction, for each $\ell \in [L-1]$ there exists a candidate bent hyperplane v_ℓ arising from layer ℓ and a candidate bent hyperplane $v_{\ell+1}$ arising from layer $\ell+1$ such that $(v_\ell, v_{\ell+1}) \in E_{f_\theta}$. Hence the dependency graph contains a chain $v_1 \to v_2 \to \cdots \to v_L$. where each v_ℓ is associated with a neuron in hidden layer ℓ.

Suppose there exists a generic parameter η in an architecture $\mathcal{A}'$ with $L' < L$ hidden layers such that $f_\eta = f_\theta$. According to Proposition 3.25, there must exist a map $\phi_\eta \colon V_{f_\theta} \to V_{\mathcal{A}'}$ such that, for every edge $(u, v) \in E_{f_\theta}$, the corresponding hidden layers satisfy $\ell_{\phi(u)} < \ell_{\phi(v)}$. Applying this property to the chain in our dependency graph, the assigned layer indices in the architecture $\mathcal{A}'$ must satisfy: $\ell_{\phi(v_1)} < \cdots < \ell_{\phi(v_L)}$. This condition requires at least L distinct layer indices. However, an architecture with only $L' < L$ hidden



(a) The function (b) One network computing the function (c) Another network computing the function (d) Another network computing the function

Figure 5: Canonical complexes for Example 6.2. Dashed lines indicate breakpoints that are not visible from the function alone, i.e., faces of the canonical complex that are not part of the breakpoint complex. Red denotes breakpoints from the second layer, and black from the first.

layers provides at most L' available layer indices. Since a strictly increasing sequence of L integers cannot be contained in the set $\{1, \ldots, L'\}$, there is no generic parameter in $\Theta_{\mathcal{A}'}$ that computes f_θ. The chain $v_1 \to \cdots \to v_L$ in the dependency graph arises from transverse visible intersections between adjacent layers in the constructed parameter. These intersections persist under sufficiently small perturbations, and hence the same chain exists for all parameters in some open neighborhood U of θ. Therefore the above obstruction applies to every parameter in U. $\square$

Without the genericity assumption, it is no longer guaranteed that one can determine the layer ordering of the neurons associated to the different facets of the breakpoint complex of f_θ, as illustrated by the following example.

Example 6.2. Consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by $x \mapsto \min\{0, \max\{x_2 - x_1 + 1, -x_2\}\}$, whose breakpoint complex $\mathcal{B}_f$ consists of three rays emanating from the vertex $(1, 0)$, as shown in Figure 5a.

For each of these three rays, there exists a parameter η realizing f in such a way that the chosen ray originates from a first-layer neuron, while the other two rays originate from the second hidden layer. This demonstrates that the internal hierarchy of the network is not a functional invariant for non-generic parameters. The polyhedral subdivisions for these three different realizations are shown in Figure 5b–Figure 5d, and the respective weights and biases are given below.

1. The first layer hyperplanes are the axes $x_1 = 0$ and $x_2 = 0$

$$W^{(1)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, b^{(1)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad W^{(2)} = \begin{pmatrix} 1 & -1 \\ 1 & -2 \end{pmatrix}, b^{(2)} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad W^{(3)} = \begin{pmatrix} -1 & 1 \end{pmatrix}.$$

2. The first layer hyperplanes are $x_1 - x_2 - 1 = 0$ and $x_1 + x_2 = 0$

$$W^{(1)} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, b^{(1)} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad W^{(2)} = \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & -3/2 \end{pmatrix}, b^{(2)} = \begin{pmatrix} -1/2 \\ -1/2 \end{pmatrix}, \quad W^{(3)} = \begin{pmatrix} -1 & 1 \end{pmatrix}.$$

3. The first layer hyperplanes are $x_2 + 1 = 0$ and $-x_1 + 2x_2 + 1 = 0$

$$W^{(1)} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}, b^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad W^{(2)} = \begin{pmatrix} 1 & -1 \end{pmatrix}, b^{(2)} = -1, \quad W^{(3)} = -1.$$

Consequently, for non-generic parameters, the geometry of the breakpoint complex alone is insufficient to identify the layer from which a specific non-linearity originates.

7 Open Questions

The polyhedral framework introduced in this work establishes the existence of identifiable parameters and minimal but non-identifiable parameters for deep ReLU networks, and describes the local functional invariants. However, several important questions remain open for future research:

1. **Measure of the Identifiable Set:** While we have shown that, for architectures with non-output layers of width at least 2, both the sets of identifiable parameters and non-identifiable parameters are open, the volume and distribution of the identifiable set are not yet well understood.

   Such an investigation will likely require identifying conditions beyond the relatively strong TPIC that still guarantee identifiability. In particular, characterizing the boundary at which identifiability fails is an important open problem. The problem is of interest for both strict identifiability and finite identifiability.

2. **Characterizing Fibers and Redundancies in Deep Networks:** While our results identify open regions of identifiable parameters and open regions with positive-dimensional fibers, the global structure of the fibers of the realization map remains poorly understood. A major open problem is to characterize, as explicitly as possible, the set of parameters representing a given function, including both its continuous and discrete redundancies. This includes finding useful semi-algebraic descriptions of fibers, understanding how visible breakpoint geometry constrains hidden cancellations, and classifying the nontrivial symmetries that can persist in minimal deep ReLU networks. A first step in this direction is given in Appendix C, where we describe an algebraic superset of the generic fiber obtained by fixing the combinatorial assignment of candidate bent hyperplanes and activation patterns.

3. **Identifiability for Width 1:** Our results apply to architectures with non-output layers of width at least 2. As discussed in Remark 4.14, if an architecture contains layers of width 1, the activation becomes one-dimensional, making it impossible to satisfy TPIC in subsequent layers. It remains an open problem whether such architectures have any identifiable parameters, or whether identifiability can hold at best in a finite sense.

4. **Identifiability Over Finite Data:** We investigated identifiability of the parameters in relation to the represented functions. Also of interest is the investigation of identifiability when one only has access to certain measurements taken from the functions, such as the function values on some finite input data set, or a list of moments of the outputs over an input data distribution.

Acknowledgments

We would like to thank Elisenda Grigsby for insightful discussions on the depth hierarchy for generic parameters and for bringing to our attention the example of Ramakrishnan (2026) showing that TPIC and LRA are not sufficient for identifiability. This project has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project 464109215 within the priority programme SPP 2298 "Theoretical Foundations of Deep Learning". GM was partially supported by DARPA AIQ grant HR00112520014, NSF grants DMS-2522495, DMS-2145630, CCF-2212520, and BMFTR in DAAD project 57616814 (SECAI).

References

Mary Phuong and Christoph H. Lampert. Functional vs. parametric equivalence of ReLU networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Bylx-TNKvH.

Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden symmetries of ReLU networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11734–11760. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/grigsby23a.html.

David Rolnick and Konrad Kording. Reverse-engineering deep ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8178–8187. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rolnick20a.html.

Héctor J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992. URL https://doi.org/10.1016/S0893-6080(05)80037-1.

Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13, 05 2000a. URL https://doi.org/10.1016/S0893-6080(00)00009-5.

Henning Petzka, Martin Trimmel, and Cristian Sminchisescu. Notes on the symmetries of 2-layer ReLU-networks. In *NLDL*, pages 1–6, 2020. URL https://doi.org/10.7557/18.5150.

Vahid Shahverdi, Giovanni Luca Marchetti, Georg Bökman, and Kathlén Kohn. Identifiable equivariant networks are layerwise equivariant, 2026a. URL https://arxiv.org/abs/2601.21645.

Věra Kůrková and Paul C. Kainen. Functionally equivalent feedforward neural networks. *Neural Computation*, 6(3):543–558, 1994. URL https://doi.org/10.1162/neco.1994.6.3.543.

Charles Fefferman. Reconstructing a neural net from its output. *Revista Matematica Iberoamericana*, 10: 507–555, 1994. URL https://api.semanticscholar.org/CorpusID:121350232.

Verner Vlačić and Helmut Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021. URL https://www.sciencedirect.com/science/article/pii/S0001870820305132.

K. Fukumizu and S. Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000b. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(00)00009-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000095.

Berfin Simsek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning*, pages 9722–9732. PMLR, 2021.

Marco Nurisso, Pierrick Leroy, Giovanni Petri, and Francesco Vaccarino. Topology and geometry of the learning space of ReLU networks: connectivity and singularities. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=O4Oy7NsSwG.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces, 2025. URL https://arxiv.org/abs/2506.13018.

J. Alexander and A. Hirschowitz. Polynomial interpolation in several variables. *Journal of Algebraic Geometry*, 4(2):201–222, 1995. MR 1311347 (96f:14065).

Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. Learning on a razor's edge: Identifiability and singularity of polynomial neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=L5jYWeycAx.

Konstantin Usevich, Ricardo Augusto Borsoi, Clara Dérand, and Marianne Clausel. Identifiability of deep polynomial neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=MrUsZfQ9pC.

Bella Finkel, Jose Israel Rodriguez, Chenxi Wu, and Thomas Yahl. Activation degree thresholds and expressiveness of polynomial neural networks. 2025. URL https://arxiv.org/abs/2408.04569.

Hoang V. Tran, Thieu Vo, An Nguyen The, Tho Tran Huu, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Minh Nguyen. Equivariant neural functional networks for transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uBai0ukstY.

Guido Montúfar. Restricted Boltzmann machines: Introduction and review. In *Information Geometry and Its Applications*, pages 75–115, Cham, 2018. Springer International Publishing. URL https://doi.org/10.1007/978-3-319-97798-0_4.

Maria Angelica Cueto, Jason Morton, and Bernd Sturmfels. Geometry of the Restricted Boltzmann Machine. *Algebraic Methods in Statistics and Probability II*, 31:135–153, 2010. doi: 10.1090/conm/516/10134. URL https://arxiv.org/abs/0908.4425.

Guido Montúfar and Jason Morton. Dimension of marginals of Kronecker product models. *SIAM Journal on Applied Algebra and Geometry*, 1(1):126–151, 2017. URL https://doi.org/10.1137/16M1077489.

Pranavkrishnan Ramakrishnan. A complete symmetry classification of shallow ReLU networks, 2026. URL https://arxiv.org/abs/2604.14037.

Steffen Dereich and Sebastian Kassing. On minimal representations of shallow ReLU networks. *Neural Networks*, 148:121–128, 2022. URL https://www.sciencedirect.com/science/article/pii/S0893608022000065.

Steffen Dereich and Sebastian Kassing. On the existence of optimal shallow feedforward networks with ReLU activation. *Journal of Machine Learning*, 3(1):1–22, 2024. URL http://dx.doi.org/10.4208/jml.230903.

Joachim Bona-Pellissier, François Bachoc, and François Malgouyres. Parameter identifiability of a deep feedforward ReLU neural network. *Machine Learning*, 112(11):4431–4493, 2023. URL https://doi.org/10.1007/s10994-023-06355-4.

Pierre Stock and Rémi Gribonval. An embedding of relu networks and an analysis of their identifiability. *Constructive Approximation*, 57(2):853–899, 2023. URL https://doi.org/10.1007/s00365-022-09578-1.

J. Elisenda Grigsby, Kathryn Lindsey, Robert Meyerhoff, and Chenxi Wu. Functional dimension of feedforward ReLU neural networks. *Advances in Mathematics*, 482:110636, 2025. URL https://doi.org/10.1016/j.aim.2025.110636.

J. Elisenda Grigsby and Kathryn Lindsey. On functional dimension and persistent pseudodimension, 2024. URL https://arxiv.org/abs/2410.17191.

Yulia Alexandr and Guido Montúfar. Constraining the outputs of ReLU neural networks, 2025. URL https://arxiv.org/abs/2508.03867.

Quynh Nguyen, Marco Mondelli, and Guido Montúfar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8119–8129. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nguyen21g.html.

Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=x8DNliTBSYY.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022. doi: 10.1214/22-AOS2211.

Kedar Karhadkar, Michael Murray, and Guido Montúfar. Bounds for the smallest eigenvalue of the NTK for arbitrary spherical data of arbitrary dimension. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=mHVmsy9len.

Kedar Karhadkar, Michael Murray, Hanna Tseran, and Guido Montúfar. Mildly overparameterized ReLU networks have a favorable loss landscape. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=10WARaIwFn.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. volume 49 of *Proceedings of Machine Learning Research*, pages 907–940, Columbia University, New York, New York, USA, 2016. PMLR. URL http://proceedings.mlr.press/v49/eldan16.html.

Matus Telgarsky. benefits of depth in neural networks. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, Columbia University, New York, New York, USA, 2016. PMLR. URL https://proceedings.mlr.press/v49/telgarsky16.html.

Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. When and why are deep networks better than shallow ones? In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 2343–2349. AAAI Press, 2017. doi: 10.1609/aaai.v31i1.10913.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/fa6f2a469cc4d61a92d96e74617c3d2a-Paper.pdf.

Razvan Pascanu, Guido Montúfar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. In *International Conference on Learning Representations*, 2014. URL https://openreview.net/forum?id=bSaT4mmQt84Lx.

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854. PMLR, August 2017. URL https://proceedings.mlr.press/v70/raghu17a.html.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4558–4566. PMLR, July 2018. URL https://proceedings.mlr.press/v80/serra18b.html.

Randall Balestriero, Romain Cosentino, Behnaam Aazhang, and Richard Baraniuk. The geometry of deep networks: Power diagram subdivision. In *Advances in Neural Information Processing Systems 32*, pages 15832–15841. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9712-the-geometry-of-deep-networks-power-diagram-subdivision.pdf.

Ekin Ergen and Moritz Grillo. Topological expressivity of relu neural networks. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1599–1642. PMLR, 2024. URL https://proceedings.mlr.press/v247/ergen24a.html.

Guido Montúfar, Yue Ren, and Leon Zhang. Sharp bounds for the number of regions of maxout networks and vertices of Minkowski sums. *SIAM Journal on Applied Algebra and Geometry*, 6(4):618–649, 2022. URL https://doi.org/10.1137/21M1413699.

Andrei Balakin, Shelby Cox, Georg Loho, and Bernd Sturmfels. Maxout polytopes, 2025. URL https://arxiv.org/abs/2509.21286.

Christoph Hertrich, Amitabh Basu, Marco Di Summa, and Martin Skutella. Towards lower bounds on the depth of ReLU neural networks. *SIAM Journal on Discrete Mathematics*, 37(2):997–1029, 2023. URL https://doi.org/10.1137/22M1489332.

Christian Alexander Haase, Christoph Hertrich, and Georg Loho. Lower bounds on the depth of integral ReLU neural networks via lattice polytopes. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=2mvALOAWaxY.

Gennadiy Averkov, Christopher Hojny, and Maximilian Merkert. On the expressiveness of rational ReLU neural networks with bounded depth. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uREg3OHjLL.

Moritz Leo Grillo, Christoph Hertrich, and Georg Loho. Depth-bounds for neural networks via the braid arrangement. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=XO9fhSZkBh.

Egor Bakaev, Florestan Brunck, Christoph Hertrich, Jack Stade, and Amir Yehudayoff. Better neural network expressivity: Subdividing the simplex, 2026. URL https://arxiv.org/abs/2505.14338.

Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5824–5832, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/zhang18i.html.

Vasileios Charisopoulos and Petros Maragos. A tropical approach to neural networks with piecewise linear activations, 2018. URL https://arxiv.org/abs/1805.08749.

Marie-Charlotte Brandenburg, Georg Loho, and Guido Montúfar. The real tropical geometry of neural networks for binary classification. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=I7JWf8XA2w.

Joey Huchette, Gonzalo Muñoz, Thiago Serra, and Calvin Tsay. When deep learning meets polyhedral theory: A survey, 2023. URL https://arxiv.org/abs/2305.00241.

Ngoc M. Tran and Jidong Wang. Minimal representations of tropical rational functions. *Algebraic Statistics*, 15(1):27–59, May 2024. URL http://dx.doi.org/10.2140/astat.2024.15.27.

Marie-Charlotte Brandenburg, Moritz Leo Grillo, and Christoph Hertrich. Decomposition polyhedra of piecewise linear functions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vVCHWVBsLH.

Diane Maclagan and Bernd Sturmfels. *Introduction to Tropical Geometry*. Graduate Studies in Mathematics. 2015.

Marissa Masden. Algorithmic determination of the combinatorial structure of the linear regions of ReLU neural networks. *SIAM Journal on Applied Algebra and Geometry*, 9(2):374–404, 2025. URL https://doi.org/10.1137/24M1646996.

J. Elisenda Grigsby and Kathryn Lindsey. On transversality of bent hyperplane arrangements and the topological expressiveness of ReLU neural networks. *SIAM Journal on Applied Algebra and Geometry*, 6 (2):216–242, 2022. URL https://doi.org/10.1137/20M1368902.

Itay Safran. A depth hierarchy for computing the maximum in ReLU networks via extremal graph theory, 2026. URL https://arxiv.org/abs/2601.01417.

A Remarks on Previous Identifiability Results and Constructions

This appendix clarifies two points relevant to comparison with prior work: first, the distinction between uniqueness within restricted parameter classes and identifiability in the full parameter space; second, the way in which our construction differs from previous reverse-engineering-based constructions.

A.1 On restricted notions of identifiability in prior work

We briefly clarify a point about the interpretation of previous identifiability results for deep ReLU networks. There are several different notions of uniqueness that one may consider. Given a parameter θ, one may ask whether $f_\eta = f_\theta$ implies $\eta \sim \theta$

1. among all parameters η,
2. among all generic parameters η, or
3. only among parameters η belonging to some restricted class, for example parameters satisfying geometric conditions such as TPIC and LRA.

These are progressively weaker notions of uniqueness and should not be conflated. In this hierarchy, the result of Rolnick and Kording (2020) is of the third type, being formulated on a restricted class of parameters satisfying geometric assumptions such as TPIC and LRA, while the result of Phuong and Lampert (2020) is of the second type, establishing uniqueness among generic parameters. That argument of Rolnick and Kording (2020) reconstructs parameters under assumptions such as TPIC and LRA, and therefore yields uniqueness only within the class of parameters satisfying those assumptions. Technically, this does not imply identifiability among all generic parameters, since a second generic realization of the same function need not a priori satisfy TPIC and LRA; nor does it imply identifiability in the full parameter space. In particular, uniqueness inside the TPIC/LRA class does not exclude the possibility of additional realizations outside that class.

This subtlety also affects how one interprets later results that build on the reverse-engineering framework such as the work of Grigsby et al. (2023). For example, if one constructs a parameter satisfying TPIC and LRA and then invokes the reverse-engineering theorem, the conclusion obtained directly is uniqueness only among other parameters satisfying the same conditions. An additional argument is required to transfer this to uniqueness among all generic parameters, and a further argument is required to upgrade this to identifiability among all parameters. In our proof, the first upgrade is achieved using the breakpoint complex and dependency graph, which show that cTPIC and LRA are inherited across generic realizations of the same function; the second upgrade is achieved by a dimension argument excluding the image of the nongeneric locus on a nonempty open subset.

A.2 Comparison of our construction to prior constructions

Our construction of identifiable parameters for deep ReLU networks is inspired by the construction of Grigsby et al. (2023). There are, however, important differences between the two approaches.

We begin by recalling the main ideas of their construction, which proceeds iteratively, layer by layer. A key ingredient is the notion of a *positive-axis hyperplane*, namely a hyperplane in the image space of the preceding layer whose defining affine function has strictly positive coefficients and sufficiently large negative bias. Such hyperplanes intersect every ray in the image of a ReLU layer and are used to enforce intersections with the image of the hyperplanes within a distinguished unbounded polyhedron. The construction maintains, at each step, a distinguished unbounded polyhedron in the canonical polyhedral complex whose image under the truncated network remains full-dimensional in the positive orthant. Subsequent layers are chosen by perturbing a high-bias positive-axis hyperplane so that its pullback intersects all bent hyperplanes from the previous layer. This iterative procedure is designed to enforce pairwise transverse intersections between bent hyperplanes of subsequent layers, enabling the application of the reverse-engineering framework of Rolnick and Kording (2020).

Our construction is similar in that it also proceeds iteratively and designs each layer to enforce (c)TPIC and LRA. However, instead of working with positive-axis hyperplanes and unbounded polyhedra, we work with bounded polytopes and slab layers. Moreover, we incorporate transparency together with cancellation-freeness to ensure that the relevant intersections remain visible in the final realized function.

A key difference is that the inductive construction of Grigsby et al. (2023) does not appear to guarantee the linear regions assumption (LRA) in a neighborhood of the relevant intersections for the final realized function. While their construction ensures pairwise transverse intersections in the canonical bent-hyperplane arrangement, this does not by itself imply that these intersections remain visible as breakpoints of the realized function. Indeed, later layers may be inactive in a neighborhood of such intersections, so that the affine linear maps on the adjacent polyhedra of the canonical complex coincide in the final output (and are in fact constant). In that case, the intersection exists in the canonical complex but is absent from the breakpoint complex, and LRA fails near that point.

This phenomenon is illustrated in Figure 6 for the architecture $(2, 2, 2, 2)$. In the left panel, the pullback of a positive-axis hyperplane along the first layer produces a bent hyperplane in the input space. In the middle panel, perturbing this hyperplane yields two second-layer bent hyperplanes. As seen in the figure, only one orientation of these hyperplanes is compatible with the inductive choice of an unbounded polyhedron S_2. In the right panel, the same procedure is applied to the third layer. Although the blue bent hyperplanes intersect those from earlier layers in the canonical complex, the third layer is inactive near the point τ, so this point is not a breakpoint of the final function. Consequently, the local four-region configuration used in Lemmas D.11 and D.14 in the work of Grigsby et al., 2023 need not correspond to four distinct linear regions of the realized function. In particular, an intersection between two bent hyperplanes from the first two hidden layers is not visible as breakpoint in the final function.

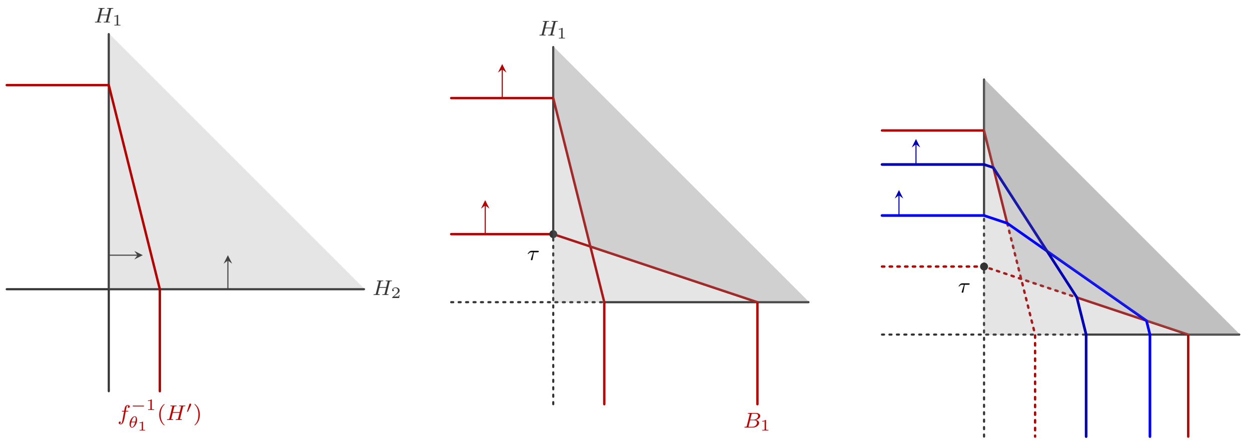
In our framework, this issue is addressed by explicitly distinguishing between the canonical polyhedral complex and the breakpoint complex. The missing ingredient in the earlier construction is a mechanism that prevents later layers from canceling previous visible nonlinearities. To address this, our proof of Theorem 4.10 imposes transparency and leverages generic cancellation-freeness to ensure LRA via the structure of tropical weights (Proposition 3.6).

B Observations on Minimal Layer Width

The following one-hidden-layer lemmas collected in this section observe the effect of "linear" neurons on minimality: neurons that do not induce visible breakpoint hyperplanes may still be necessary. Once the visible nonlinear structure is fixed, the remaining discrepancy between two realizations is forced to be affine, and the rank of this affine part can require additional hidden neurons that are invisible from the breakpoint geometry alone. In this sense, affine contributions can enforce minimal width even when they do not create any new visible breakpoints.

Lemma B.1. *Let $P \subseteq \mathbb{R}^d$ be a polytope, and let $f_\theta(x) = W^{(2)}[W^{(1)}x + b^{(1)}]_+ + b^{(2)}$ be a one-hidden-layer network with n hidden neurons such that the breakpoint set of f_θ in $\operatorname{relint}(P)$ is the union of n pairwise distinct hyperplanes $H_i = \{x \in \mathbb{R}^d \mid W_i^{(1)}x + b_i^{(1)} = 0\}$, for $i \in [n]$. Let f_η be another function realized by a one-hidden-layer network with n hidden neurons whose breakpoint hyperplanes intersecting $\operatorname{relint}(P)$ are the same hyperplanes $H_1, \dots, H_n$. If $f_\theta - f_\eta$ is affine linear on P, then the linear part of $f_\theta - f_\eta$ on P is of the form $\sum_{i=1}^n \alpha_i W_{:,i}^{(2)} W_i^{(1)}$ with $\alpha_i \in \{-1, 0, 1\}$.*

Proof. Since both networks have exactly n neurons and exactly the same n pairwise distinct breakpoint hyperplanes in $\operatorname{relint}(P)$, after reordering the neurons of f_η we may assume that for each $i \in [n]$, the ith neuron of f_η has breakpoint hyperplane H_i. Let $f_\eta(x) = V^{(2)}[V^{(1)}x + c^{(1)}]_+ + c^{(2)}$. For each i, the ith neuron of f_η therefore has preactivation of the form $V_i^{(1)}x + c_i^{(1)} = \lambda_i(W_i^{(1)}x + b_i^{(1)})$ for some nonzero scalar $\lambda_i \in \mathbb{R}$. Since $[V_i^{(1)}x + c_i^{(1)}]_+$ has breakpoint hyperplane H_i, the sign of λ_i determines on which side of H_i the neuron is active.



(a) The first hidden layer consists of the coordinate-axis hyperplanes H_1 and H_2. The pullback of a positive-axis hyperplane H' in the image is a bent hyperplane in the input space.

(b) After perturbation, the second-layer bent hyperplanes are shown in red. The dark gray region is the distinguished unbounded polyhedron S_2 used in the inductive construction.

(c) The same step is repeated for the third layer (blue). Although the new bent hyperplanes intersect the earlier ones in the canonical complex, the point τ is no longer visible as a breakpoint of the final function.

Figure 6: Illustration of the inductive construction of Grigsby et al. (2023) for the architecture $(2, 2, 2, 2)$. Black lines are the first-layer hyperplanes, red lines the second-layer bent hyperplanes, and blue lines the third-layer bent hyperplanes. The shaded regions indicate the distinguished unbounded polyhedra used in the construction. Dashed segments represent pieces of bent hyperplanes that are present in the canonical complex but do not survive as breakpoints of the final function.

By Proposition 3.13, the tropical weight of f_θ along H_i is $c_{f_\theta}(H_i) = W^{(2)}_{:,i}\|W^{(1)}_i\|$. Likewise, the tropical weight of f_η along H_i is $c_{f_\eta}(H_i) = V^{(2)}_{:,i}\|V^{(1)}_i\|$. Since $f_\theta - f_\eta$ is affine linear on P, its tropical weight vanishes on every facet contained in H_i, hence $c_{f_\theta}(H_i) = c_{f_\eta}(H_i)$ for all $i \in [n]$. Therefore $V^{(2)}_{:,i}\|V^{(1)}_i\| = W^{(2)}_{:,i}\|W^{(1)}_i\|$. Since $V^{(1)}_i = \lambda_i W^{(1)}_i$, we have $\|V^{(1)}_i\| = |\lambda_i|\,\|W^{(1)}_i\|$, and hence $V^{(2)}_{:,i} = \frac{1}{|\lambda_i|}W^{(2)}_{:,i}$.

Now consider the contribution of the i-th neuron to the linear part on a region R of $\mathcal{C}_\theta(P)$. If $\lambda_i > 0$, then the ith neuron of f_η is active on exactly the same side of H_i as the ith neuron of f_θ, and its linear contribution on that side is $V^{(2)}_{:,i}V^{(1)}_i = \frac{1}{|\lambda_i|}W^{(2)}_{:,i} \cdot \lambda_i W^{(1)}_i = W^{(2)}_{:,i}W^{(1)}_i$. If $\lambda_i < 0$, then it is active on the opposite side, and its linear contribution there is $V^{(2)}_{:,i}V^{(1)}_i = -W^{(2)}_{:,i}W^{(1)}_i$.

Hence, the linear part of f_θ on R is $A^\theta_R = \sum_{i \in S_\theta(R)} W^{(2)}_{:,i}W^{(1)}_i$, while the linear part of f_η on R is $A^\eta_R = \sum_{i \in S_\eta(R)} \pm W^{(2)}_{:,i}W^{(1)}_i$, where for each i, the sign is determined by λ_i. Therefore the difference $A^\theta_R - A^\eta_R$ is of the form $\sum_{i=1}^n \alpha_i W^{(2)}_{:,i}W^{(1)}_i$ with $\alpha_i \in \{-1, 0, 1\}$. Since $f_\theta - f_\eta$ is affine linear on P, this difference is independent of the choice of region R, and equals the linear part of $f_\theta - f_\eta$ on P. This proves the claim. □

Lemma B.2. *Let $P \subseteq \mathbb{R}^d$ be a polytope, and let $f(x) = V^{(2)}[V^{(1)}x + c^{(1)}]_+ + c^{(2)}$ be a one-hidden-layer network with $n + k$ hidden neurons. Assume that the breakpoint set of f in* relint(P) *is the union of exactly n pairwise distinct hyperplanes $H_1, \ldots, H_n$. Then there exists another one-hidden-layer network $\tilde{f}(x) = \tilde{V}^{(2)}[\tilde{V}^{(1)}x + \tilde{c}^{(1)}]_+ + \tilde{c}^{(2)}$ with $n + k$ hidden neurons such that:*

1. *$\tilde{f}(x) = f(x)$ for all $x \in P$;*
2. *exactly n neurons of $\tilde{f}$ have breakpoint hyperplanes intersecting* relint(P)*, namely $H_1, \ldots, H_n$;*

3. the remaining k neurons have breakpoint hyperplanes disjoint from P.

Proof. Group the neurons of f according to the visible hyperplane they induce in $\mathrm{relint}(P)$. For each $j \in [n]$, let I_j be the set of neurons whose breakpoint hyperplane in $\mathrm{relint}(P)$ is H_j, and let K be the set of neurons whose nonzero locus does not appear as breakpoint of the final function, that is, whose breakpoint hyperplanes do not intersect $\mathrm{relint}(P)$ or are canceled by some other neurons breakpoint hyperplane. Then $[n+k] = I_1 \sqcup \cdots \sqcup I_n \sqcup K$.

Fix for each $j \in [n]$ an affine linear map $h_j(x) = u_j x + \beta_j$ with zero set H_j. Every neuron in I_j has preactivation of the form $\lambda_{j,\ell} h_j(x)$ for some $\lambda_{j,\ell} \neq 0$. Using $[-h_j(x)]_+ = [h_j(x)]_+ - h_j(x)$, we can rewrite the total contribution of the neurons in I_j as $\sum_{\ell \in I_j} V^{(2)}_{:,\ell}[\lambda_{j,\ell} h_j(x)]_+ = a_j[h_j(x)]_+ + b_j h_j(x)$ for suitable vectors $a_j, b_j \in \mathbb{R}^m$.

Thus $f(x) = g(x) + \alpha(x)$ for all $x \in P$, where $g(x) \coloneqq \sum_{j=1}^n a_j[h_j(x)]_+$ has exactly n visible neurons with breakpoint hyperplanes $H_1, \ldots, H_n$, and α is an affine linear map.

By construction, g has the same tropical weight as f along each visible hyperplane H_j. Since f and g are compatible with the same arrangement on P, Lemma 2.1 implies that $f - g = \alpha$ is affine linear on P.

It remains to realize the affine linear map α by k neurons whose hyperplanes are disjoint from P. Replacing the $|I_j|$ neurons on H_j by one visible neuron removes $\sum_{j=1}^n (|I_j| - 1)$ neurons, and together with the neurons in K this leaves $|K| + \sum_{j=1}^n (|I_j| - 1) = (n+k) - n = k$ neurons available for the affine part.

The affine linear map α is the sum of the contributions of exactly these k leftover neurons on P, so its linear part has rank at most k. Conversely, any affine linear map of rank r can be represented on P by r neurons whose hyperplanes are disjoint from P, together with the output bias term. Hence α can be represented on P by at most k such neurons; if fewer are needed, we add dummy neurons with zero output weights.

Combining these k outside neurons with the visible network g yields a one-hidden-layer network $\tilde{f}$ with $n+k$ hidden neurons that agrees with f on P, has exactly n visible neurons with breakpoint hyperplanes $H_1, \ldots, H_n$, and whose remaining k neurons have breakpoint hyperplanes disjoint from P. ☐

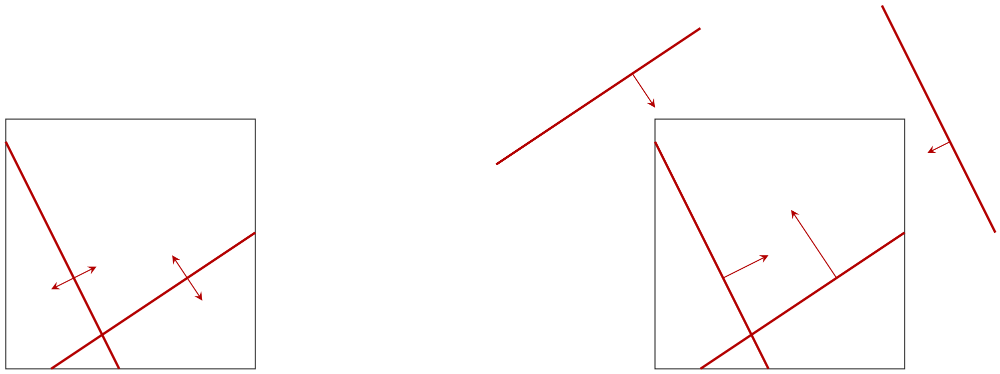
See also Figure 7 for an illustration.

Lemma B.3. *Let $P \subseteq \mathbb{R}^d$ be a polytope of dimension at least r, and let Q denote the orthogonal projection onto $\mathrm{aff}(P)$. Let $f_\theta \colon \mathbb{R}^d \to \mathbb{R}^n \to \mathbb{R}^m$ be a generic one-hidden-layer network with parameter $\theta = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ such that all hyperplanes of $\mathcal{H}_{W^{(1)},b^{(1)}}$ are inside P. Let $A \in \mathbb{R}^{m \times d}$ be a matrix such that $(A + \sum_{i=1}^n \alpha_i W^{(2)}_{:,i} W^{(1)}_i)Q$ has rank r for all $\alpha_i \in \{-1, 0, 1\}$. Then any one-hidden-layer network f satisfying $f(x) = f_\theta(x) + Ax$ for all $x \in P$ must have at least $n + r$ hidden neurons.*

Proof. Let $f(x) = V^{(2)}[V^{(1)}x + c^{(1)}]_+ + c^{(2)}$ be a one-hidden-layer network with N hidden neurons such that $f(x) = f_\theta(x) + Ax$ for all $x \in P$. Since Ax is affine linear, the breakpoint set of f in $\mathrm{relint}(P)$ coincides with that of f_θ. By assumption, f_θ has exactly n distinct breakpoint hyperplanes $H_1, \ldots, H_n$ inside P. By Lemma B.2, there exists another one-hidden-layer network $\tilde{f}$ with N hidden neurons such that:

1. $\tilde{f}(x) = f(x)$ for all $x \in P$;
2. exactly n neurons of $\tilde{f}$ have breakpoint hyperplanes intersecting $\mathrm{relint}(P)$, namely $H_1, \ldots, H_n$;
3. the remaining $N - n$ neurons have breakpoint hyperplanes disjoint from P.

Write $\tilde{f}(x) = g(x) + h(x)$ for all $x \in P$, where g is the subnetwork consisting of the n visible neurons and h is the subnetwork consisting of the remaining $N - n$ neurons. Then h is affine linear on P, since all of its breakpoint hyperplanes are disjoint from P. Moreover, $g(x) = f_\theta(x) + Ax - h(x)$ for all $x \in P$, so $f_\theta - g$ is affine linear on P. Since both f_θ and g have exactly n visible neurons with visible hyperplanes $H_1, \ldots, H_n$, Lemma B.1 implies that the linear part of $f_\theta - g$ is of the form $\sum_{i=1}^n \alpha_i W^{(2)}_{:,i} W^{(1)}_i$. Hence the affine-linear map $h(x) = Ax + (f_\theta - g)(x)$ has linear part $A + \sum_{i=1}^n \alpha_i W^{(2)}_{:,i} W^{(1)}_i$. By assumption, the projection $(A + \sum_{i=1}^n \alpha_i W^{(2)}_{:,i} W^{(1)}_i)Q$ has rank r. Therefore the linear part of $h|_{\mathrm{aff}(P)}$ has rank at least r.



(a) Several neurons induce the same visible breakpoint hyperplanes inside P.

(b) Only one neuron remains on each visible breakpoint hyperplane, whereas the remaining affine linear function is computed by neurons whose hyperplanes lie outside of P and hence are affine linear on P.

Figure 7: Illustration of Lemma B.2.

On the other hand, h is represented by $N - n$ neurons, each of which is affine linear on P and contributes a matrix of rank at most 1 to the linear part. It follows that $N - n \geq r$. Therefore $N \geq n + r$, which proves the claim. □

C Algebraic Superset of Fiber

In this section, we describe (a superset of) the fiber algebraically. To do so, we first fix the combinatorial structure of the parameterization and then deriving the necessary polynomial constraints. Specifically, we first choose a discrete assignment of candidate bent hyperplanes to and an activation pattern for each facet, and then we establish equations for the weights and biases based on two local functional invariants: the affine span of the facets (local geometry) and the gradient jumps across them (tropical weights). The resulting set is a superset of the fiber, as it captures the necessary local algebraic consistency conditions but ignores the global semi-algebraic sign constraints and the base affine map required for functional identity.

Definition C.1 (Discrete fiber configuration)**.** Let $\theta \in \Theta_{\mathcal{A}_{\mathrm{orig}}}$ be a generic parameter, and let f_θ be the corresponding network function. For each facet $\sigma \in \mathcal{B}_\theta^{d-1}$ of the breakpoint complex, let $c_\theta(\sigma)$ be the corresponding non-zero tropical weight, and let $a_\sigma \in \mathbb{R}^d$ be a unit vector and $\beta_\sigma \in \mathbb{R}$ be such that $\sigma \subseteq \{x \in \mathbb{R}^d \mid \langle a_\sigma, x\rangle + \beta_\sigma = 0\}$.

Given another (possibly different) architecture $\mathcal{A}$, a *discrete fiber configuration* for f_θ with respect to $\mathcal{A}$ is a pair $(\phi, \mathbf{s})$, where:

1. $\phi\colon V_{f_\theta} \to V_{\mathcal{A}}$ is a map from the set of candidate bent hyperplanes of f_θ to the set of hidden neurons of $\mathcal{A}$ such that for every edge $(u, v) \in E_{f_\theta}$, if $\phi(u) = (i, \ell)$ and $\phi(v) = (j, k)$, then $\ell < k$;
2. $\mathbf{s} = \{\mathbf{s}(\sigma)\}_{\sigma \in \mathcal{B}_\theta^{d-1}}$ is an assignment of an activation pattern to each facet $\sigma \in \mathcal{B}_\theta^{d-1}$.

For such a configuration, we write $\varphi\colon \mathcal{B}_\theta^{d-1} \to V_{\mathcal{A}}$ with $\varphi(\sigma) \coloneqq \phi(\pi(\sigma))$, where $\pi\colon \mathcal{B}_\theta^{d-1} \to V_{f_\theta}$ maps a facet to its unique candidate bent hyperplane.

Definition C.2. Let $(\phi, \mathbf{s})$ be a discrete fiber configuration for f_θ with respect to $\mathcal{A}$, and let $\varphi(\sigma) = \phi(\pi(\sigma))$ as above. We say that a parameter $\eta \in \Theta_{\mathcal{A}}$ *realizes* $(\phi, \mathbf{s})$ if, for every facet $\sigma \in \mathcal{B}_\theta^{d-1}$,

1. the facet σ is contained in the bent hyperplane of the neuron $\varphi(\sigma)$ in the realization η, and

2. the activation pattern induced by η on σ is equal to $\mathbf{s}(\sigma)$.

Definition C.3 (Configuration variety)**.** For an architecture $\mathcal{A}$ and a discrete fiber configuration $(\phi, \mathbf{s})$ for f_θ, let $\varphi(\sigma) = \phi(\pi(\sigma))$ for all $\sigma \in \mathcal{B}_\theta^{d-1}$. The *configuration variety* $\mathcal{V}_{(\phi,\mathbf{s})} \subseteq \Theta_\mathcal{A} \times \mathbb{R}^{\mathcal{B}_\theta^{d-1}} \times \mathbb{R}^{\mathcal{B}_\theta^{d-1}}$ is the algebraic variety in the variables $\eta = (W^{(\ell)}, b^{(\ell)})_{\ell\in[L+1]}$ and $(\lambda_\sigma, \delta_\sigma)_{\sigma\in\mathcal{B}_\theta^{d-1}}$ defined as follows.

For each facet $\sigma \in \mathcal{B}_\theta^{d-1}$, let $\varphi(\sigma) = (j, \ell)$ and let $S_k(\sigma) = \{r \in [n_k] \mid \mathbf{s}_{k,r}(\sigma) = +\}$. We write $g_\sigma(\eta) \coloneqq \left(W^{(\ell)} D_{S_{\ell-1}(\sigma)} W^{(\ell-1)} \cdots D_{S_1(\sigma)} W^{(1)}\right)_j \in \mathbb{R}^d$ for the j-th row of the matrix product, i.e., the gradient of the (j, ℓ)-neuron preactivation evaluated under the activation pattern $\mathbf{s}(\sigma)$, and

$$t_\sigma(\eta) \coloneqq \left(b^{(\ell)} + \sum_{k=1}^{\ell-1} W^{(\ell)} D_{S_{\ell-1}(\sigma)} W^{(\ell-1)} \cdots D_{S_{k+1}(\sigma)} W^{(k+1)} D_{S_k(\sigma)} b^{(k)}\right)_j$$

for the value of this preactivation at the origin. Moreover, we define

$$v_\sigma(\eta) \coloneqq W^{(L+1)} D_{S_L(\sigma)} W^{(L)} \cdots D_{S_{\ell+1}(\sigma)} W^{(\ell+1)} e_j.$$

Then $\mathcal{V}_{(\phi,\mathbf{s})}$ is cut out by the following equations, imposed for every facet $\sigma \in \mathcal{B}_\theta^{d-1}$:

1. Geometric Alignment: $g_\sigma(\eta) = \delta_\sigma \lambda_\sigma a_\sigma$ and $t_\sigma(\eta) = \delta_\sigma \lambda_\sigma \beta_\sigma$.

2. Tropical Weight Matching: $\lambda_\sigma v_\sigma(\eta) = c_\theta(\sigma)$.

3. Sign Equation: $\delta_\sigma^2 = 1$.

For fixed activation patterns $\mathbf{s}(\sigma)$, all expressions $g_\sigma(\eta)$, $t_\sigma(\eta)$, and $v_\sigma(\eta)$ are polynomial in the parameters η. Hence the above equations define an algebraic variety. We denote by $\pi_\mathcal{A}(\mathcal{V}_{(\phi,\mathbf{s})})$ the projection onto $\Theta_\mathcal{A}$ and call this the *configuration set*. For a CPWL function f and an architecture $\mathcal{A}$, let $\tilde{\mathcal{S}}(f, \mathcal{A}) = \{\eta \in \tilde{\Theta}_\mathcal{A} \mid f = f_\eta\}$ be its generic fiber.

Proposition C.4. *Let θ be a generic parameter and $\mathcal{A}$ a (potentially different) architecture. For any discrete fiber configuration $(\phi, \mathbf{s})$ for f_θ with respect to $\mathcal{A}$, let $\tilde{\mathcal{S}}(f_\theta, \phi, \mathbf{s}) \coloneqq \{\eta \in \tilde{\mathcal{S}}(f_\theta, \mathcal{A}) \mid \eta$ realizes $(\phi, \mathbf{s})\}$. Then $\tilde{\mathcal{S}}(f_\theta, \phi, \mathbf{s}) \subseteq \pi_\mathcal{A}(\mathcal{V}_{(\phi,\mathbf{s})})$. Consequently, the generic fiber $\tilde{\mathcal{S}}(\theta)$ is contained in the finite union $\tilde{\mathcal{S}}(\theta) \subseteq \bigcup_{(\phi,\mathbf{s})} \pi_\mathcal{A}(\mathcal{V}_{(\phi,\mathbf{s})})$, taken over all valid discrete fiber configurations.*

Proof. Let $\eta \in \tilde{\mathcal{S}}(f_\theta, \phi, \mathbf{s})$. Then $f_\eta = f_\theta$, and η realizes the configuration $(\phi, \mathbf{s})$.

Fix a facet $\sigma \in \mathcal{B}_\theta^{d-1}$ and let $\varphi(\sigma) = \phi(\pi(\sigma)) = (j, \ell)$. Since η realizes $(\phi, \mathbf{s})$, the facet σ is contained in the bent hyperplane of neuron (j, ℓ), and the activation pattern induced by η on σ is equal to $\mathbf{s}(\sigma)$. Hence the corresponding preactivation vanishes on σ. Therefore, setting $\lambda_\sigma \coloneqq \|g_\sigma(\eta)\|_2 \geq 0$, there exists $\delta_\sigma \in \{-1, 1\}$ such that $g_\sigma(\eta) = \delta_\sigma \lambda_\sigma a_\sigma$ and $t_\sigma(\eta) = \delta_\sigma \lambda_\sigma \beta_\sigma$. Since $f_\eta = f_\theta$, the tropical weights agree on the breakpoint complex: $c_\eta(\sigma) = c_\theta(\sigma)$. On the other hand, by Proposition 3.6, the tropical weight induced by the neuron (j, ℓ) under the activation pattern $\mathbf{s}(\sigma)$ is given by $c_\eta(\sigma) = \lambda_\sigma v_\sigma(\eta)$. Hence $\lambda_\sigma v_\sigma(\eta) = c_\theta(\sigma)$.

Since this holds for every facet $\sigma \in \mathcal{B}_\theta^{d-1}$, we obtain $(\eta, (\lambda_\sigma, \delta_\sigma)_\sigma) \in \mathcal{V}_{(\phi,\mathbf{s})}$. Therefore $\eta \in \pi_\mathcal{A}(\mathcal{V}_{(\phi,\mathbf{s})})$, which proves the claim. □