The Symmetries of Three-Layer ReLU Networks

Johanna Marie Gegenfurtner
Technical University of Denmark
johge@dtu.dk

Moritz Grillo
Max Planck Institute for Mathematics in the Sciences
Leipzig
moritz.grillo@mis.mpg.de

Guido Montúfar
UCLA and MPI MiS
montufar@math.ucla.edu

Abstract

We develop a framework for analyzing parameter symmetries in deep ReLU networks and obtain a complete characterization of the generic parameter fibers for three-layer bottleneck architectures. Our approach provides explicit semi-algebraic descriptions of these fibers and yields a polynomial time algorithm for deciding functional equivalence of two parameters. The symmetries include discrete and continuous transformations arising from layer composition, and depend on whether deeper layers hide or preserve geometric structure from preceding layers. Finally, we show that some of these symmetries induce local conservation laws along gradient flow, while others do not.

1 Introduction

A fundamental property of neural network parameterizations is parameter redundancy: distinct parameter configurations can realize the same function. These redundancies can be formalized via the notion of *fibers*, defined as the set of all parameters realizing a given function. Symmetries, i.e., parameter transformations that leave the function unchanged, provide a canonical mechanism for generating such fibers. For instance, in ReLU networks, neuron weights can be permuted within layers, and incoming and outgoing weights rescaled without affecting the function (Kůrková and Kainen, 1994, Rolnick and Kording, 2020). Beyond such global symmetries, fibers also arise from *subnetwork* structure: when the realized function depends only on a subset of parameters, the rest can vary freely without changing the function. Such symmetries significantly influence the optimization landscape and can induce conserved quantities along training, constraining how gradient methods explore parameter space and which solutions are ultimately selected (Głuch and Urbanke, 2021, Kunin et al., 2021, Zhao et al., 2023, 2026, Nurisso et al., 2026).

In ReLU networks, parameter symmetries are relatively well understood in the single-hidden-layer case. In particular, Petzka et al. (2020) show that, for generic parameters, the fiber is trivial up to permutation and scaling, and provide a description in the remaining degenerate cases. In deep architectures, the situation is more nuanced. While global symmetries are still generally limited to permutation and scaling, additional local symmetries can arise at specific non-degenerate parameter configurations (Rolnick and Kording, 2020, Grigsby et al., 2023). At the same time, for many ReLU architectures there exist functions whose parameters are identifiable (i.e., uniquely determined) among suitable classes of parameters up to these global symmetries (Phuong and Lampert, 2020, Grigsby et al., 2023). However, identifiability in ReLU networks is inherently non-uniform and the degree of redundancy can vary significantly across parameter space (Grigsby et al., 2025). In particular, the existence of identifiable parameter configurations does not imply identifiability throughout generic parameters.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

As a result, beyond well-understood global symmetries, the structure of fibers in ReLU networks with more than one hidden layer remains largely uncharacterized. In particular, a systematic description of all parameters realizing a given function, and of the interaction between different sources of redundancy, is still missing. Addressing this problem for two-hidden-layer networks is a key step toward understanding identifiability, minimality, and parameter optimization in deep ReLU networks.

1.1 Our Contributions

We develop a systematic framework for describing symmetries and fibers in ReLU neural networks beyond the standard scaling and permutation symmetries. Our contributions are as follows:

1. **Layerwise symmetries:** We characterize symmetries acting within individual layers by analyzing two-layer networks on the nonnegative orthant. In this setting, we obtain explicit descriptions of fibers via polynomial equations and inequalities and compute the dimension of generic fibers. These results provide a fundamental building block for the analysis of deeper architectures.
2. **Symmetries induced by layer composition:** We identify new symmetries that emerge from composing ReLU layers by analyzing three-layer networks. We show that layer composition introduces redundancies that cannot be reduced to symmetries acting within individual layers.
3. **Explicit fiber description for three-layer architectures:** We show the above symmetries are complete for three-layer bottleneck networks with generic parameters. Remarkably, these fibers depend only on weight signs and imply a polynomial time algorithm for deciding functional equivalence of parameters. An implementation of the algorithm can be found under `https://github.com/MoGrillo/relu-bottleneck-equivalence`.
4. **Non-localizability of fibers:** We show that symmetries in deep networks are not always localizable. We construct networks where each pair of consecutive layers is identifiable, but their composition is not, showing that parameter redundancy can arise as a global property of network composition.
5. **Connections to training dynamics:** We show that local symmetries from linearly acting neurons induce conserved quantities along gradient flow, but the symmetries from Section 5 do not.

1.2 Related Work

The study of parameter equivalence in neural networks dates back at least to the work of Sussmann (1992), Kůrková and Kainen (1994), Fefferman (1994) for tanh and asymptotically constant activations. For networks with polynomial activations, recent works show, depending on the degree, that generic parameters are identifiable up to permutation and scaling (Shahverdi et al., 2026, Usevich et al., 2025, Finkel et al., 2025). More generally, Vlačić and Bölcskei (2021) develop a framework in which parameter fibers are generated by affine symmetries of the activation function, leading to equivalence classes beyond permutation and scaling. For ReLU activations, however, they indicate that additional, non-affine symmetries would need to be taken into account to characterize the fibers.

For ReLU networks, parameter symmetries are particularly intricate. In the shallow setting, Petzka et al. (2020), Ramakrishnan (2026) give a complete classification, showing that the generic fiber is trivial up to permutation and scaling, and characterizing degenerate cases where subsets of neurons aggregate into a single linear function. In a related direction, Dereich and Kassing (2022) characterize minimal representations of scalar-valued functions realized by shallow ReLU networks and determine the dimension and number of connected components of the corresponding function space. Building on this, Dereich and Kassing (2024) prove the existence of optimal approximations of continuous target functions within this function space. Their tools are closely related to the hyperplane representation used in our work, which generalizes this viewpoint to intermediate layers with nonnegative inputs and to non-scalar outputs.

Our approach complements the reverse-engineering framework of Rolnick and Kording (2020), which leverages the combinatorial geometry of activation boundaries to recover parameters. Relatedly, Grigsby et al. (2023) show that architectures with widths at least the input dimension admit open sets of parameters that are identifiable within a suitable parameter class, while Phuong and Lampert (2020) prove analogous results for architectures with non-increasing widths. However, the structure of the fibers outside these identifiable sets remains largely unknown.

We further highlight the work of Stock and Gribonval (2023), which introduces a path-based locally linear parametrization of the realization map and derives conditions for local identifiability. Com-

plementary local analyses by Bona-Pellissier et al. (2022, 2023) provide conditions under which deep ReLU networks are identifiable. From a different perspective, Elbrächter et al. (2019) study the degeneracy of the ReLU parametrization via inverse stability, showing the realization map can be highly non-injective. To our knowledge, a general description of parameter fibers in ReLU networks with more than one hidden layer, particularly accounting for compositional effects, remains unknown.

Our discussion of training dynamics is motivated by the well-known fact that parameter symmetries can induce conservation laws during training. For example, matrix product symmetries preserve imbalance quantities (Tarmoun et al., 2021, Kunin et al., 2021), a line of work further developed by Le and Jegelka (2022), Marcotte et al. (2023). These conserved quantities are of interest because they constrain the set of reachable parameters and provide insight into implicit bias and generalization. Recent work further characterizes initialization-dependent invariant sets of gradient flow and shows that their singularities correspond to submodels (Nurisso et al., 2026). This raises the question of which additional conserved quantities may arise from other symmetry mechanisms.

2 Preliminaries

For $n \in \mathbb{N}$, we write $[n] \coloneqq \{1, \dots, n\}$. For $x \in \mathbb{R}^d$, we denote by $[x]_+$ the entrywise application of the ReLU activation function. For a set of indices $S \subseteq [n]$, let $D_S \in \mathbb{R}^{n \times n}$ denote the diagonal matrix with $(D_S)_{ii} = 1$ if $i \in S$ and $(D_S)_{ii} = 0$ otherwise. We write $\| \cdot \|$ for the Euclidean norm.

ReLU networks. A *ReLU layer* with $n_{\ell-1}$ inputs and n_ℓ outputs is the map $\phi^{(\ell)} \colon \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$; $x \mapsto [W^{(\ell)}x + b^{(\ell)}]_+$, parametrized by weight matrix $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ and bias vector $b^{(\ell)} \in \mathbb{R}^{n_\ell}$. A feedforward *ReLU network* with architecture $\mathcal{A} = (n_0, \dots, n_{L+1})$ is the composition $f_\theta(x) = T^{(L+1)} \circ \phi^{(L)} \circ \cdots \circ \phi^{(1)}(x)$, where the final layer $T^{(L+1)}(y) = W^{(L+1)}y + b^{(L+1)}$ is affine. We write $d = n_0$ and $m = n_{L+1}$ for the input and output dimensions. The parameter $\theta = (W^{(1)}, b^{(1)}, \dots, W^{(L+1)}, b^{(L+1)})$ lies in the *parameter space* $\Theta_\mathcal{A} \cong \bigoplus_{\ell=1}^{L+1} \left(\mathbb{R}^{n_\ell \times n_{\ell-1}} \times \mathbb{R}^{n_\ell}\right)$. The associated *realization map* is $\mu_\mathcal{A} \colon \Theta_\mathcal{A} \to \mathcal{F}_\mathcal{A}$; $\theta \mapsto f_\theta$. For $\ell \in [L]$, define the *preactivation* and *activation* at layer ℓ by $z^{(\ell,\theta)}(x) \coloneqq W^{(\ell)}a^{(\ell-1,\theta)}(x) + b^{(\ell)}$ and $a^{(\ell,\theta)}(x) \coloneqq [z^{(\ell,\theta)}(x)]_+$, with $a^{(0,\theta)}(x) = x$. For fixed θ, the maps f_θ, $z^{(\ell,\theta)}$, and $a^{(\ell,\theta)}$ are continuous and piecewise linear. For a parameter $\theta \in \Theta_\mathcal{A}$, we also use $W^{(\theta,\ell)}$ and $b^{(\theta,\ell)}$ to denote the weight matrix and bias vector of the ℓ-th layer, but omit θ in the superscript when there is no risk of ambiguity.

Fibers and symmetries. The *fiber* of a parameter $\theta \in \Theta_\mathcal{A}$ is $\mathcal{S}(\theta) \coloneqq \{\eta \in \Theta_\mathcal{A} \mid f_\eta = f_\theta\}$. The parameter space of ReLU networks carries a standard equivalence relation generated by:

1. permutation of neurons within any hidden layer, obtained by permuting the rows of $[W^{(\ell)}, b^{(\ell)}]$, and inversely permuting the columns of $W^{(\ell+1)}$;
2. positive rescaling of a hidden neuron, obtained by multiplying the corresponding row of $[W^{(\ell)}, b^{(\ell)}]$ by $\lambda > 0$, and multiplying the corresponding column of $W^{(\ell+1)}$ by λ^{-1}.

We write $\theta \sim \eta$ if η is obtained from θ by a sequence of such transformations. A parameter θ is called *identifiable* if $f_\eta = f_\theta$ implies $\eta \sim \theta$.

Polyhedral complexes and piecewise linear functions. A *polyhedral complex* $\mathcal{C}$ in $\mathbb{R}^d$ is a finite collection of polyhedra containing the empty set such that if $P \in \mathcal{C}$, then every face of P belongs to $\mathcal{C}$, and if $P, Q \in \mathcal{C}$ with $P \cap Q \neq \emptyset$, then $P \cap Q$ is a face of both. We write $\mathcal{C}^k$ for the k-faces of $\mathcal{C}$, and call $\mathcal{C}^{d-1}$ the *facets* and $\mathcal{C}^d$ the *regions*. For a hyperplane arrangement $\mathcal{H}$, we write $\mathcal{C}_\mathcal{H}$ for the induced polyhedral complex. A function $f \colon \mathbb{R}^d \to \mathbb{R}^m$ is *continuous piecewise linear* (CPWL) if there exists a complete polyhedral complex $\mathcal{C}$ such that $f|_P$ is affine for each $P \in \mathcal{C}$; in this case, f and $\mathcal{C}$ are *compatible*. A point $x \in \mathbb{R}^d$ is a *breakpoint* of f if no neighborhood of x exists on which f is affine. We denote the set of breakpoints by $B(f)$. For further details, see Appendix A.

Weighted complexes. Let $f \colon \mathbb{R}^d \to \mathbb{R}^m$ be CPWL and 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 adjacent maximal cells with $P \cap Q = \sigma$, and suppose $f(x) = A_P x + b_P$ for $x \in P$ and $f(x) = A_Q x + b_Q$ for $x \in Q$. If $e_{P/\sigma}$ denotes the unit normal to σ pointing from Q into P, define the *weight* of f along σ by $c_f(\sigma) \coloneqq (A_P - A_Q)e_{P/\sigma} \in \mathbb{R}^m$. The

map $c_f : \mathcal{C}^{d-1} \to \mathbb{R}^m$ records the *gradient jump* of f across facets. The following standard fact from polyhedral and tropical geometry will be used repeatedly. More details are provided in Appendix A.

Theorem 1 (Maclagan and Sturmfels, 2015)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be CPWL and compatible with a polyhedral complex $\mathcal{C}$. Then the weight function $c_f : \mathcal{C}^{d-1} \to \mathbb{R}^m$ uniquely determines f up to a global affine linear function.*

3 Polyhedral Geometry of ReLU Networks

In this section, we recall polyhedral structures naturally associated with ReLU networks that will be needed for the fiber analysis. Additional background and proofs are deferred to Appendix B.

Canonical polyhedral complex and bent hyperplanes. For a ReLU network f_θ with architecture $\mathcal{A} = (n_0, n_1, \ldots, n_L, n_{L+1})$, the *canonical polyhedral complex* $\mathcal{C}_\theta$ is constructed iteratively: starting from $\mathcal{C}_{\theta,0} = \mathbb{R}^d$, one subdivides each cell of $\mathcal{C}_{\theta,\ell-1}$ by the pulled back hyperplanes induced by the preactivations of layer ℓ. The resulting complex $\mathcal{C}_\theta = \mathcal{C}_{\theta,L}$ is indexed by global activation patterns, and f_θ restricts to an affine-linear map on every maximal cell. For a neuron (j, ℓ), the preactivation $z_j^{(\ell,\theta)}$ is affine linear on each cell $R \in \mathcal{C}_{\theta,\ell-1}$. On cells where it is non-constant, its zero set is a codimension-one polyhedron and the union of these polyhedra is called the neuron's *bent hyperplane*, denoted $B_{\ell,j}(\theta)$. We denote by $\mathcal{H}_\ell(\theta) = \{B_{\ell,j}(\theta) \mid j \in [n_\ell]\}$ the set of bent hyperplanes of neurons in layer ℓ. We note that not every facet of the canonical polyhedral complex is necessarily associated to a change in the linear behavior of the realized function. The *breakpoint complex* is the subcomplex $\mathcal{B}_\theta \coloneqq \{P \in \mathcal{C}_\theta \mid P \subseteq B(f_\theta)\}$, whose support is the breakpoint set: $|\mathcal{B}_\theta| = B(f_\theta)$. The canonical complex $\mathcal{C}_\theta$ carries a weight function c_θ induced by f_θ, as in Theorem 1. A facet of $\mathcal{C}_\theta$ is visible in the breakpoint complex precisely when the corresponding weight is nonzero: $\mathcal{B}_\theta^{d-1} = \{\sigma \in \mathcal{C}_\theta^{d-1} \mid c_\theta(\sigma) \neq 0\}$.

Generic parameters. Many of our results concern *generic* parameters. At an informal level, genericity combines a geometric and an algebraic requirement. Geometrically, the bent hyperplanes induced by the network intersect in the expected codimensions, so that the canonical polyhedral complex has the expected combinatorics. Algebraically, all masked products of weight matrices $W^{(\ell)} D_{S_{\ell-1}} W^{(\ell-1)} \cdots D_{S_k} W^{(k)}$ have maximal possible rank, preventing degeneracies in the resulting linear parts, such as accidental alignments or cancellations. The set of generic parameters is open and dense in $\Theta_\mathcal{A}$; we denote it by $\tilde{\Theta}_\mathcal{A} \subseteq \Theta_\mathcal{A}$. We refer to Appendix B.2 for a precise formulation.

Visibility of first-layer hyperplanes in three-layer networks. For three-layer networks, a first-layer hyperplane may or may not remain visible in the realized function, depending on the activity of the second layer. The key mechanism is whether the hyperplane is *anchored*, meaning that it is intersected by, or lies in the active region of, a second-layer bent hyperplane. The following result is illustrated in Figure 1 and proved in Appendix B.6. It will be central in the bottleneck analysis.

Theorem 2. *Let $\mathcal{A} = (d, n_1, n_2, n_3)$ with $d, n_1 > 1$, and let $\theta, \eta \in \Theta_\mathcal{A}$ be generic parameters with $f_\theta = f_\eta$. Let $\mathcal{H}_1(\theta)$ denote the set of first-layer hyperplanes of θ. Then:*

1. *If $H \in \mathcal{H}_1(\theta)$ and there exists a neuron $j \in [n_2]$ together with a point $x \in H$ such that $z_j^{(2,\theta)}(x) > 0$, then $H \in \mathcal{H}_1(\eta)$.*
2. *In particular, if this holds for every $H \in \mathcal{H}_1(\theta)$, then $\mathcal{H}_1(\theta) = \mathcal{H}_1(\eta)$ and $\mathcal{C}_\theta = \mathcal{C}_\eta$.*

This theorem shows that first-layer hyperplanes that are functionally visible through the second layer are rigid along the fiber. The only possible source of nontrivial fiber arising from the first layer in the three-layer case is therefore the presence of *hidden hyperplanes*, which we analyze in Section 5.

4 Layerwise Fibers

In this section, we provide a complete description of the fiber for a one-hidden-layer ReLU network $f_\theta : \mathbb{R}^d_{\geq 0} \to \mathbb{R}^m$. This is essential for understanding deep networks, as the output of any ReLU layer (except the final one) serves as a non-negative input to the subsequent layer. We first establish a

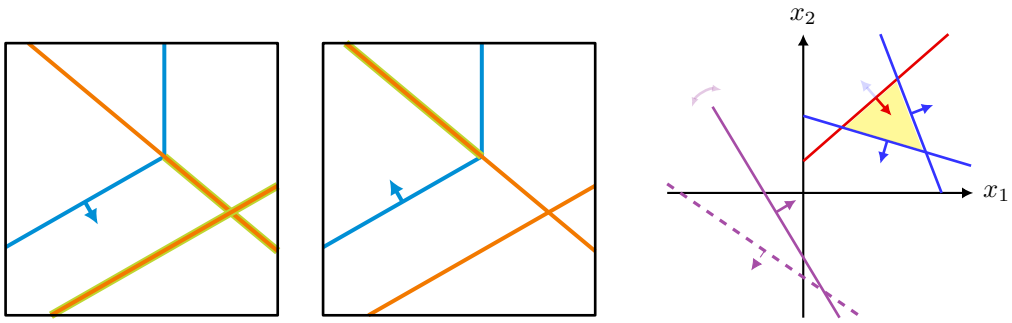

Figure 1: **Left:** Illustration of Theorem 2. The orange hyperplanes correspond to first-layer neurons, and the cyan bent hyperplane to a second-layer neuron. The portions of the first-layer hyperplanes that intersect the active region of the second-layer neuron are highlighted in green. **Right:** Illustration of sign-flipping mechanism. The region R is shown in yellow. Hyperplanes of nonlinear neurons that are active on R are shown in red, and those that are inactive in blue. A dead neuron is shown as a dashed, and a linear neuron as a solid purple plane. If the orientation of nonlinear neurons is reversed, this must be compensated by a corresponding change in the linear neurons, as described in (1).

hyperplane representation of f_θ. We then describe the fiber as a union of semi-algebraic sets formed by intersections of algebraic varieties and polyhedral cones. Finally, we compute the dimension of the generic fiber components. All proofs belonging to this section can be found in Appendix C.

4.1 Hyperplane Representation

Let $X \subseteq \mathbb{R}^d$ be a full-dimensional polyhedron; in this paper we will only consider the cases $X = \mathbb{R}^d_{\geq 0}$ or $X = \mathbb{R}^d$. A one-hidden-layer ReLU network $f_\theta \colon X \to \mathbb{R}^m$ is a CPWL function whose breakpoints are supported on a hyperplane arrangement $\mathcal{H} = \{H_1, \ldots, H_k\}$. By Proposition 22 in the appendix, the weight function c_θ is constant along each breakpoint hyperplane: if $\sigma, \sigma' \in \mathcal{C}_\theta^{d-1}$ are facets contained in the same hyperplane H_i, then $c_\theta(\sigma) = c_\theta(\sigma')$. Thus each breakpoint hyperplane H_i carries a well-defined vector weight $c(i) \coloneqq c_\theta(\sigma)$ for any facet $\sigma \subseteq H_i$.

By Theorem 1, the weighted hyperplane arrangement $(\mathcal{H}, c)$ determines f_θ up to the addition of a global affine-linear function. This remaining ambiguity is removed by fixing the affine map on a full-dimensional region R of the arrangement. This motivates the following definition.

Definition 3. A tuple $(\mathcal{H}, A_R, b_R, c)$ is called a *hyperplane representation* of $f_\theta \colon X \to \mathbb{R}^m$ if:

1. $\mathcal{H} = \{H_1, \ldots, H_k\}$ is the arrangement covering the breakpoints of f_θ in X;
2. $R \in \mathcal{C}_\mathcal{H} \cap X$ is a full-dimensional region and $f_\theta(x) = A_R x + b_R$ for all $x \in R$;
3. $c \colon [k] \to \mathbb{R}^m$ satisfies $c(i) = c_\theta(\sigma)$ for every facet $\sigma \in \mathcal{C}_\theta^{d-1}$ with $\sigma \subseteq H_i$.

The following lemma shows that hyperplane representations completely characterize one-hidden-layer functions on the orthant.

Lemma 4. *For every one-hidden-layer network $f_\theta \colon \mathbb{R}^d_{\geq 0} \to \mathbb{R}^m$, there exists a hyperplane representation $(\mathcal{H}, A_R, b_R, c)$. Moreover, this representation fully characterizes the function: for any parameter η, we have $f_\eta = f_\theta$ if and only if $(\mathcal{H}, A_R, b_R, c)$ is also a hyperplane representation of f_η.*

4.2 Explicit Description of the Fibers

By Lemma 4, a one-hidden-layer network f_θ is completely determined by its hyperplane representation $(\mathcal{H}, A_R, b_R, c)$. Hence, to describe the fiber $\mathcal{S}(\theta)$, it suffices to characterize all parameters that realize this same hyperplane representation. In full generality, this leads to the semi-algebraic description given in Appendix C.2. Here, we restrict attention to generic and minimal parameters, where that description simplifies considerably.

We call a hidden neuron i *linear (on $\mathbb{R}^d_{\geq 0}$)* if its preactivation is nonnegative on $\mathbb{R}^d_{\geq 0}$, equivalently, in the one-hidden-layer case, if $W_i^{(1)} \geq 0$ and $b_i^{(1)} \geq 0$. Let $(\mathcal{H}, A_R, b_R, c)$ be the hyperplane

representation of f_θ, and write $\mathcal{H} = \{H_1, \dots, H_k\}$. For each $i \in [k]$, choose $a_i \in \mathbb{R}^d$ and $t_i \in \mathbb{R}$ such that $\|a_i\| = 1$ and $H_i = \{x \in \mathbb{R}^d \mid \langle a_i, x\rangle + t_i = 0\}$. We also fix a reference region R of the arrangement and choose the signs of the a_i so that $R = \{x \in X \mid \langle a_i, x\rangle + t_i \le 0 \text{ for all } i \in [k]\}$.

For a generic parameter, the hyperplane arrangement is generic. This implies that every nonlinear neuron corresponds to a unique hyperplane in $\mathcal{H}$ and that no cancellations occur. If the parameter is moreover minimal, then there are no dead neurons. Hence the hidden neurons split into exactly two classes: the k nonlinear neurons corresponding to the breakpoint hyperplanes, and the remaining neurons, which are linear on $\mathbb{R}^d_{\ge 0}$ We denote the latter index set by $J \coloneqq [n] \setminus [k]$.

Thus, in the generic minimal case, the fiber is governed by two types of choices. First, each nonlinear neuron may realize its associated hyperplane with either orientation, giving rise to a discrete family. Second, once these orientations are fixed, the resulting affine-linear discrepancy on the reference region R must be realized by the linear neurons, producing a continuous family of factorizations.

To make this explicit, let $o \in \{1, -1\}^k$ be an orientation choice for the nonlinear neurons, and define $S \coloneqq \{i \in [k] \mid o(i) = -1\}$, so that S is the set of nonlinear neurons active on the reference region R. We write S_θ for the set of nonlinear neurons that are active on R for the original parameter θ. In the generic case, the gradient jump across H_i is induced by a single neuron, so after normalizing so that $\|W^{(1)}_i\|_2 = 1$, we have $c(i) = W^{(2)}_{:,i}$. If one changes the orientation of the nonlinear neurons from S_θ to S, then the linear neurons must compensate for the resulting change in the affine map on R. This is illustrated in Figure 1 (right panel). The required contribution from the linear neurons is given by

$$A_S \coloneqq A_R + \sum_{i \in S} c(i) a_i^\top, \qquad b_S \coloneqq b_R + \sum_{i \in S} c(i) t_i. \tag{1}$$

since $\sum_{i \in S} -c(i)a_i^\top$ is the linear contribution of the nonlinear neurons that are active on R. The semi-algebraic families introduced below naturally describe subsets of the full fiber $\mathcal{S}(\theta)$. However, they need not consist entirely of generic parameters in the sense of Definition 19. Accordingly, the generic fiber $\tilde{\mathcal{S}}(\theta) = \mathcal{S}(\theta) \cap \tilde{\Theta}_{\mathcal{A}}$ is obtained by intersecting these families with $\tilde{\Theta}_{\mathcal{A}}$.

Proposition 5. *Let θ be a generic and minimal parameter. Then the generic fiber is given by the disjoint union $\tilde{\mathcal{S}}(\theta)/\sim \;= \bigcup_{S \subseteq [k]} \left(K_S \cap \tilde{\Theta}_{\mathcal{A}}\right) / \sim$, where the union is taken over orientations $S \subseteq [k]$, and $K_S = \{p_S\} \times V_S$ is the semi-algebraic set defined as follows:*

1. *p_S: The parameters of the k nonlinear neurons are fixed as follows. Let $o \in \{1, -1\}^k$ be the orientation determined by $S = \{i \in [k] \mid o(i) = -1\}$. Then*

$$(W^{(1)}_i, b^{(1)}_i) = o(i)(a_i, t_i), \quad W^{(2)}_{:,i} = c(i) \quad \forall i \in [k]. \tag{2}$$

2. *V_S: The parameters of the linear neurons indexed by $J = [n] \setminus [k]$ and the output bias vector $b^{(2)}$ are subject to the following semi-algebraic constraints, where A_S and b_S are defined in* (1)*:*
   - normalization and positivity*: $\|W^{(1)}_i\|_2 = 1$ and $(W^{(1)}_i, b^{(1)}_i) \ge 0$ for all $i \in J$;*
   - linear factorization*: $W^{(2)}_{:,J} W^{(1)}_J = A_S$;*
   - bias alignment*: $W^{(2)}_{:,J} b^{(1)}_J + b^{(2)} = b_S$.*

In Appendix C.4, we establish the dimension of the semi-algebraic sets V_S arising in the generic case based on the concept of the *nonnegative column rank*.

Theorem 6. *Let θ be generic and minimal. For any given orientation $S \subseteq [k]$, consider the corresponding semi-algebraic component V_S of the generic fiber defined in Proposition 5. Then:*

1. *If $n - k < \min\{d, m\}$, then $\dim(V_S) = (n-k)^2$ if and only if $S = S_\theta$, and $V_S = \emptyset$ otherwise.*
2. *If $n - k = d \le m$, then $\dim(V_S) = (n-k)^2$ for all $S \subseteq [k]$.*
3. *If $n - k = m < d$, then $\dim(V_S) = (n-k)^2$ if and only if $\operatorname{cone}(A_S) \subseteq \operatorname{col}(A_S)$ is pointed, and $V_S = \emptyset$ otherwise, where A_S is the compensation defined in* (1)*.*
4. *If $n - k = m + 1 \le d$, then $\dim(V_S) = m^2 + m + d$ for all $S \subseteq [k]$.*

5 Symmetries from Layer Composition

We now analyze how interactions between layers give rise to symmetries that do not appear at the level of individual layers, particularly when certain neurons hide the geometric features of preceding layers. Throughout this section, let $\mathcal{A} = (d, n_1, n_2, m)$. These symmetries are complete for generic weights in the bottleneck case ($n_1 \leq d$) as we show in Section 6.

This subsection formalizes the symmetries that arise when neurons in one layer hide the geometric features of a preceding layer. We show that this induces a coupled symmetry: a translation of a first-layer hyperplane can be exactly compensated by a corresponding shift in the second-layer biases.

Definition 7. Let $\theta \in \Theta_{\mathcal{A}}$. A neuron $i \in [n_2]$ *hides* hyperplane $j \in [n_1]$ if $W^{(2)}_{i,j} > 0$, $W^{(2)}_{i,k} \leq 0$ for all $k \neq j$, and $b^{(2)}_i \leq 0$. The parameter θ *hides* hyperplane j if every neuron $i \in [n_2]$ does.

The Continuous Symmetry. If the second layer hides hyperplane j, then translating the first-layer neuron j by shifting $b^{(1)}_j$, induces a change in $a^{(1)}_j$ that can be compensated by a corresponding shift in the second-layer biases $b^{(2)}$. This is because, under the hiding condition, second-layer neurons depend only on movement of the j-th hyperplane in a regime where $a^{(1)}_j$ is linear, while negative weights and biases keep them inactive elsewhere. See Figure 2 for an illustration. For a parameter $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [3]} \in \tilde{\Theta}_{\mathcal{A}}$ that hides hyperplane j, we define the set of translated parameters:

$$T_j(\theta) = \left\{ (W^{(1)}, b^{(1)} + te_j, W^{(2)}, b^{(2)} - tW^{(2)\top}_{:j}, W^{(3)}, b^{(3)}) \;\middle|\; t \geq \max_{i \in [n_2]} \frac{b^{(2)}_i}{W^{(2)}_{ij}} \right\}. \tag{3}$$

Lemma 8. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [3]} \in \tilde{\Theta}_{\mathcal{A}}$ be hiding hyperplane j. Then $T_j(\theta) \subseteq \mathcal{S}(\theta)$.*

The Discrete Symmetry. While the local translation symmetry $T_j(\theta)$ exists for any hiding parameter, we identify an additional discrete symmetry that arises specifically when the second hidden layer consists of a single neuron, i.e., $n_2 = 1$. This symmetry permits a discrete "flipping" of the first-layer hyperplanes that leaves the realized function invariant.

More precisely, for $j \in [n_1]$ and any subset $I \subseteq \{1, \ldots, n_1\} \setminus \{j\}$, define $M_I = I_{n_1} - \sum_{k \in I} e_j e_k^\top - 2\sum_{k \in I} e_k e_k^\top \in \mathbb{R}^{n_1 \times n_1}$, where I_{n_1} is the identity matrix and e_k is the k-th standard basis vector. The set $\mathcal{G}_j = \{M_I \mid I \subseteq \{1, \ldots, n_1\} \setminus \{j\}\}$ is a finite Abelian group under matrix multiplication, with group operation $M_I M_J = M_{I \triangle J}$, for $I, J \subseteq \{1, \ldots, n_1\} \setminus \{j\}$, where $I \triangle J$ denotes the symmetric difference of sets. Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [3]} \in \tilde{\Theta}_{\mathcal{A}}$ be generic and hiding hyperplane j, and suppose that $n_2 = 1$. By the scaling symmetry, we may assume WLOG that $W^{(2)}_{1k} \in \{+1, -1\}$ for all $k \in [n_1]$ and $b^{(2)}_1 \in \{+1, -1\}$. We call such a parameter *normalized*.

Proposition 9. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [3]} \in \Theta_{\mathcal{A}}$ be normalized and hiding hyperplane j. For $M_I \in \mathcal{G}_j$, let $M_I \cdot (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(3)}) = (M_I W^{(1)}, M_I b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(3)})$. Then $\mathcal{G}_j \cdot \theta \subseteq \mathcal{S}(\theta)$.*

6 Three-Layer Bottleneck Fibers

We now give a complete description of generic fibers for three-layer bottleneck architectures. Throughout, let $\mathcal{A} = (d, n_1, n_2, m)$ with $n_1 \leq d$. For such architectures, the first-layer image is generically the nonnegative orthant, so the fiber structure is governed by the action of the second layer on this. Our analysis splits into two cases. If no first-layer hyperplane is hidden, the first layer is rigid along the fiber, reducing the problem to the one-hidden-layer classification in Section 4. If a hyperplane is hidden, the translation and sign-flipping symmetries from Section 5 account for the remaining generic degrees of freedom. We show that, in the bottleneck regime, these symmetries are complete.

Non-Hiding Parameters. If every first-layer hyperplane remains functionally visible through the second layer, then the first hidden layer is fixed across the fiber up to trivial symmetries.

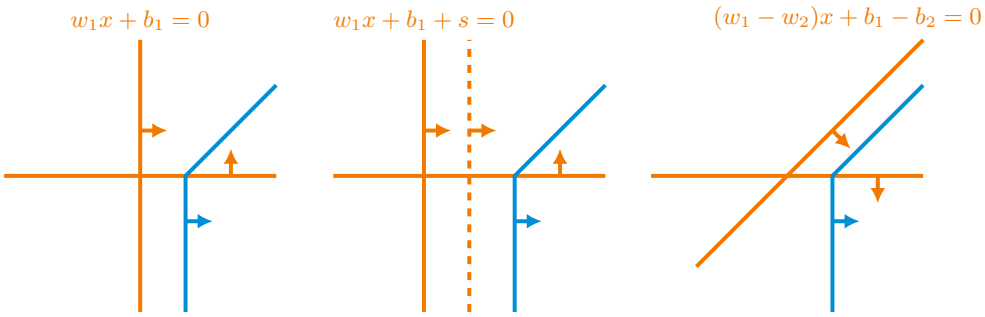

Figure 2: Parameters in the fiber when there is a hidden hyperplane. Two first-layer hyperplanes are shown in orange and one second-layer bent hyperplane is shown in blue. The left two panels illustrate the continuous symmetry and the one on the right illustrates the discrete symmetry.

Proposition 10. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ have no dead neurons and be non-hiding. Then the first-layer parameters are fixed across $\tilde{\mathcal{S}}(\theta)$ up to permutation and positive scaling, and the generic fiber reduces to the one-hidden-layer fiber of the induced map $g_{\theta'} : \mathbb{R}^{n_1}_{\geq 0} \to \mathbb{R}^m$, given by $g_{\theta'}(y) = W^{(3)}[W^{(2)}y + b^{(2)}]_+ + b^{(3)}$. In particular, $\tilde{\mathcal{S}}(\theta)/\sim \cong \tilde{\mathcal{S}}(\theta')/\sim$, where the right-hand side is described by Proposition 5.*

Proof sketch. The rigidity of the first layer follows from Theorem 2. Since $n_1 \leq d$ and $W^{(1)}$ has full rank, the first hidden layer maps surjectively onto $\mathbb{R}^{n_1}_{\geq 0}$, so equality of the realized functions reduces to equality of the induced one-hidden-layer maps on the orthant which is described in Section 4. The detailed proof is given in Appendix E.1. □

Hiding Parameters. We now turn to the case where a first-layer hyperplane is hidden by the second layer. If the second layer has width at least two, the hidden hyperplane may be translated, but its direction remains rigid.

Proposition 11. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ be generic and hiding hyperplane j, and assume that $n_2 \geq 2$. Then $\tilde{\mathcal{S}}(\theta)/\sim = (T_j(\theta) \cap \tilde{\Theta}_{\mathcal{A}})/\sim$.*

If the second hidden layer consists of a single neuron, then in addition to the translation symmetry there is a finite family of discrete sign-flipping symmetries.

Proposition 12. *Let $\theta = (W^{(\ell,\theta)}, b^{(\ell,\theta)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ be generic and normalized, assume that $n_2 = 1$, and suppose that θ hides hyperplane j. Then $\tilde{\mathcal{S}}(\theta)/\sim = \left(\bigcup_{I\subseteq[n_1]\setminus\{j\}} \left(T_j(M_I \cdot \theta) \cap \tilde{\Theta}_{\mathcal{A}}\right)\right)/\sim$.*

The proofs of both propositions are deferred to Appendix E.2. Geometrically, they rely on two facts. First, all first-layer hyperplanes except the hidden one remain visible along the fiber and are therefore fixed. Second, in the case $n_2 \geq 2$, each second-layer neuron produces, on the region where only neuron j is active in the first layer, a visible facet parallel to the hidden hyperplane. Since there are at least two such parallel facets in the breakpoint complex, their common direction determines the direction of the hidden hyperplane, so the only remaining freedom is to translate it. By contrast, when $n_2 = 1$, there is only one second-layer bent hyperplane. Its normal vectors can then be realized by several different sign patterns of the first layer, and these alternative realizations give rise to the discrete sign-flipping symmetries described by the matrices M_I.

Characterization of Identifiability. By combining non-hiding and hiding cases, we obtain a complete classification of identifiability for three-layer bottleneck networks with generic parameters based entirely on sign patterns of second-layer weights and biases.

Theorem 13. *Let $\mathcal{A} = (d, n_1, n_2, m)$ with $d \geq n_1 \geq 2$. Then a generic parameter $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ is non-identifiable if and only if at least one of the following holds:*

1. There exists a row of $[W^{(2)}, b^{(2)}]$ with all entries negative.

2. *There exists a row of* $[W^{(2)}, b^{(2)}]$ *with all entries positive.*

3. *There exists* $j \in [n_1]$ *such that* $[W^{(2)}, b^{(2)}]$ *is positive in column* j *and negative elsewhere.*

Proof. Conditions (1) produces a dead neuron, and (2) produces a neuron that acts linearly, leading to a positive-dimensional fiber by Theorem 6. Condition (3) induces the hiding symmetries T_j and $\mathcal{G}_j$ which are distinct from the permutation and scaling symmetry. If neither holds, then by Proposition 10 the first layer is rigid and the second-layer fiber is trivial, implying identifiability.

Condition (1) produces a dead neuron, thus we may replace the corresponding row by zeroes. This parameter can not be obtained from θ through permutation or rescaling. Condition (2) produces a neuron that acts linearly, leading to a positive-dimensional fiber by Theorem 6. Condition (3) induces the hiding symmetries T_j and $\mathcal{G}_j$ which are also distinct from the permutation and scaling symmetry. Thus, in either of these cases, θ is non-identifiable.

Suppose that neither 1-3 holds. Since 3 is not satisfied, by Proposition 10 the first layer hyperplanes are all visible as breakpoints in f_θ, and thus the first layer parameters are fixed up to positive rescaling and permutations. Further, since 1,2 are not satisfied, and the second-layer parameters admit no additional symmetries, implying identifiability. □

Since this yields a complete characterization and containment in the explicit semi-algebraic sets can be evaluated in polynomial time, we obtain the following.

Proposition 14. *There exists a polynomial-time algorithm that, given a bottleneck architecture* $\mathcal{A} = (d, n_1, n_2, m)$ *with* $n_1 \leq d$ *and two generic parameters* $\theta, \eta \in \tilde{\Theta}_\mathcal{A}$*, decides whether* $f_\theta = f_\eta$*.*

Algorithm 1 Deciding if $f_\theta = f_\eta$ in the bottleneck setting.

Require: generic parameters θ, η with $n_1 \leq d$.
Remove dead neurons, i.e. rows of $[W^{(2)}, b^{(2)}]$ with all entries ≤ 0.
if all second layer neurons for both θ, η are dead **then** ▷ f_θ, f_η are constant valued
    if $b^{(3,\theta)} = b^{(3,\eta)}$ **then return** TRUE
    if $b^{(3,\theta)} \neq b^{(3,\eta)}$ **then return** FALSE
else if exactly one has no active second layer neurons **then return** FALSE
From the sign pattern of $[W^{(2)}, b^{(2)}]$, decide whether θ, η are hiding, with hidden index j.
if exactly one of θ, η is hiding **then return** FALSE
if θ, η are non-hiding **then**
    if rows of $[W^{(1)}, b^{(1)}]$ do not match up to permutation and positive scaling **then return** FALSE ▷ Lemma 46
    if the weighted hyperplane arrangements of the induced networks on $\mathbb{R}^{n_1}_{\geq 0}$ agree and their affine maps agree on one base region **then return** TRUE ▷ Lemma 4
    return FALSE
if θ, η are hiding **then**
    if η does not hide the same H_j, or the non-hidden rows of $[W^{(1)}, b^{(1)}]$ do not match up to permutation and scaling **then return** FALSE
    if $n_2 = 1$ **then**
        $I \leftarrow$ indices of matched non-hidden rows with negative scaling
        $\theta \leftarrow M_I \cdot \theta$ ▷ Proposition 12
    if some $t \in \mathbb{R}$ satisfies the linear (in)equalities of Equation (3) **then return** TRUE ▷ Lemma 48
    return FALSE

The relative volume of the set of identifiable parameters can be computed from Theorem 13. Simple lower and upper bounds can be given as follows:

Corollary 15. *Let* $\mathcal{A} = (d, n_1, n_2, m)$ *with* $d \geq n_1 \geq 2$*, let* $\mathcal{I} \subseteq \Theta_\mathcal{A}$ *be the set of identifiable parameters, and* $\mathbb{B}_r$ *the radius-*r *ball in* $\Theta_\mathcal{A}$ *for an arbitrary* $r > 0$*. Then*

$$1 - n_2\, 2^{-n_1} - n_1\, 2^{-n_2 n_1} \;\leq\; \frac{\mathrm{vol}(\mathcal{I} \cap \mathbb{B}_r)}{\mathrm{vol}(\Theta_\mathcal{A} \cap \mathbb{B}_r)} \;\leq\; (1 - 2^{-(n_1+1)})^{n_2}.$$

7 Implications for Deeper Networks

Fibers are not Localizable. A natural question is whether identifiability can be reduced to layerwise identifiability. For deep polynomial networks, identifiability can indeed be established by examining pairwise compositions (Usevich et al., 2025). The following proposition shows that such a reduction is not possible for ReLU networks. Specifically, identifiability in ReLU networks can depend on interactions between the range of earlier layers and the non-linearities of later ones.

Proposition 16. *Let $\mathcal{A} = (n_0, n_1, n_2, n_3, m)$ with $n_0 \geq n_1 \geq n_2$. Then there exists $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [4]}$ such that, for the layer maps $f_\ell \colon \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$ defined by $x \mapsto \left[W^{(\ell)}x + b^{(\ell)}\right]_+$, the compositions $f_2 \circ f_1$ and $f_3 \circ f_2$ are identifiable, while $f_3 \circ f_2 \circ f_1$ is not.*

Locally Conserved Quantities under Gradient Flow. We leverage the newly found symmetries on orthants of the parameter space to identify conserved quantities under gradient flow. Marcotte et al. (2023) identify quantities conserved over the entire parameter space and show that, under suitable conditions, their characterization captures all independent conserved functions. Our goal is to identify distinct functions by relaxing the requirement of global conservation, instead seeking quantities that remain constant only on subsets of the parameter space. Indeed, the layerwise continuous symmetries in Section 4 which are induced by the monoid action of GL^+, give rise to locally conserved quantities.

Theorem 17. *Let $O \subseteq \Theta_{\mathcal{A}}$ be an orthant of parameter space where ℓ-layer neurons $J \subseteq [n_\ell]$ are linear. Then $(W^{(\ell)}_{:,J})^\top W^{(\ell)}_{:,J} - W^{(\ell-1)}_J (W^{(\ell-1)}_J)^\top - b^{(\ell-1)}_J (b^{(\ell-1)}_J)^\top$ is a conserved quantity on O.*

A proof can be found in Appendix F.2; moreover we show in Appendix F.3 that the symmetries discovered in Section 5 do not induce locally conserved quantities.

8 Conclusion, Limitations, and Outlook

We developed a framework for analyzing parameter symmetries in deep ReLU networks beyond classical scaling and permutation invariances. We provided a complete description of the symmetries and the parameter fibers for three-layer bottleneck architectures with generic parameters, including explicit formulas for their dimensions and a polynomial-time algorithm for deciding functional equivalence of parameters. Our results show that parameter redundancy in deep ReLU networks arise not only from layerwise effects but also from compositional interactions across layers. A natural next step is to investigate the topology of fibers, including the number of connected components. While the layerwise and hiding symmetries described in this paper are not specific to the generic three-layer bottleneck setting, establishing their completeness and obtaining explicit fiber descriptions beyond this regime is likely substantially more involved. In wider architectures ($n_1 > d$), later-layer hyperplanes must be analyzed relative to a union of polyhedral images rather than the non-negative orthant, suggesting an exponential blow-up in the number of relevant activation patterns. In deeper networks, equivalence cannot be reduced to pairwise layer composition (Proposition 16), and one must track how the image of an entire prefix is intersected by the hyperplanes of the subsequent layers. In nongeneric settings, the geometric rigidity underlying our analysis breaks down, leading to additional degeneracies and symmetries.

These considerations indicate that the bottleneck and genericity assumptions in our main results isolate a regime in which genuinely deep compositional symmetries already arise while still admitting explicit and efficiently checkable descriptions. This perspective is consistent with complexity-theoretic evidence suggesting that our result lie near a tractability boundary for deciding functional equivalence and for obtaining explicit descriptions of ReLU network fibers. We discuss these extensions and their relation to hardness barriers in Appendix G.

Acknowledgments

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". JMG was supported by the DFF Sapere Aude Starting Grant "GADL", and would like to thank *Women in Machine Learning* (WiML) for generously supporting her conference travel. 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

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

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, 2020. URL https://proceedings.mlr.press/v119/rolnick20a.html.

Grzegorz Głuch and Rüdiger Urbanke. Noether: The more things change, the more stay the same, 2021. URL https://arxiv.org/abs/2104.05508.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=q8qLAbQBupm.

Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9ZpciCOunFb.

Bo Zhao, Robin Walters, and Rose Yu. Symmetry in neural network parameter spaces. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=jLpWq5QY6I.

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.

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.

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, 2023. URL https://proceedings.mlr.press/v202/grigsby23a.html.

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, Robert Meyerhoff, and Chenxi Wu. Functional dimension of feedforward ReLU neural networks. *Advances in Mathematics*, 482:110636, 2025. URL https://www.sciencedirect.com/science/article/pii/S0001870825005341.

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://www.sciencedirect.com/science/article/pii/S0893608005800371.

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

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*, 2026. 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. *Algebraic Statistics*, 16(2):113–130, 2025. URL https://msp.org/astat/2025/16-2/astat-v16-n2-p02-s.pdf.

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.

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. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2022.01.006. 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. ISSN 2790-2048. doi: 10.4208/jml.230903. URL http://dx.doi.org/10.4208/jml.230903.

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.

Joachim Bona-Pellissier, Francois Malgouyres, and Francois Bachoc. Local identifiability of deep ReLU neural networks: the theory. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-3cHWtrbLYq.

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.

Dennis Maximilian Elbrächter, Julius Berner, and Philipp Grohs. How degenerate is the parametrization of neural networks with the ReLU activation function? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/04115ec378e476c56d19d827bcf8db56-Paper.pdf.

Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR, 2021. URL https://proceedings.mlr.press/v139/tarmoun21a.html.

Thien Le and Stefanie Jegelka. Training invariances and the low-rank phenomenon: beyond linear networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=XEW8CQgArno.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: conservation laws for gradient flows. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=kMueEV8Eyy.

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

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.

Moritz Grillo and Guido Montúfar. Most relu networks admit identifiable parameters, 2026. URL https://arxiv.org/abs/2605.03601.

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.

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.

P. Orlik and H. Terao. *Arrangements of Hyperplanes*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1992. URL https://doi.org/10.1007/978-3-662-02772-1_2.

Vincent Froese, Moritz Leo Grillo, Christoph Hertrich, and Moritz Stargalla. Parameterized hardness of zonotope containment and neural network verification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=y8N45EEW05.

Vincent Froese, Moritz Grillo, and Martin Skutella. Complexity of injectivity and verification of relu neural networks (extended abstract). In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 2188–2189. PMLR, 2025. URL https://proceedings.mlr.press/v291/froese25a.html.

Appendix Contents

A Background on Polyhedral Geometry **14**

B Polyhedral Geometry of ReLU Networks **15**

C Layerwise Fibers **19**

D Fibers from Layer Composition **26**

E Three-Layer Bottleneck Fibers **29**

F Implications for Deeper Networks **35**

G Discussion: Beyond the Generic Bottleneck Regime **37**

A Background on Polyhedral Geometry

For a vector $a \in \mathbb{R}^d$ and a scalar $b \in \mathbb{R}$, the *hyperplane* $H \coloneqq \{x \in \mathbb{R}^d \mid \langle a, x\rangle + b = 0\}$ subdivides $\mathbb{R}^d$ into *half-spaces* $H^+ \coloneqq \{x \in \mathbb{R}^d \mid \langle a, x\rangle + b \geq 0\}$ and $H^- \coloneqq \{x \in \mathbb{R}^d \mid \langle a, x\rangle + b \leq 0\}$. A finite set of hyperplanes $\mathcal{H} = \{H_1, \ldots, H_n\}$ is called a *hyperplane arrangement*. A hyperplane arrangement $\mathcal{H}$ in $\mathbb{R}^d$ is *generic* if the intersection of any subset of $k \leq d$ 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 $\mathcal{H}^{\text{ess}}$ 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 $\mathcal{C}^d$ the *regions*, $\mathcal{C}^{d-1}$ the *facets*, and $\mathcal{C}^{d-2}$ the *ridges* of $\mathcal{C}$. The *dimension* of a complex $\mathcal{C}$ is the maximal dimension of its polyhedra. 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 within the affine space $\operatorname{aff}(\sigma)$.

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 $\tau \subseteq \sigma$, we 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 pair $(\mathcal{C}, c)$ forms a *balanced (weighted) polyhedral complex* if the weight function satisfies the *balancing condition* at every $\tau \in \mathcal{C}^{d-2}$:

$$\sum_{\sigma\in\operatorname{star}(\tau)^{d-1}} c(\sigma) \cdot e_{\sigma/\tau}^\top = 0.$$

If $\mathcal{C}$ is a polyhedral fan in $\mathbb{R}^2$, then its unique 0-dimensional face is the origin. The balancing condition then requires that, for each $i \in [m]$, the weighted sum of the unit generators of the rays in $\mathcal{C}$, $\sum_{\sigma \in \mathcal{C}^1} c_i(\sigma) \cdot e_\sigma$, equals zero. In higher dimensions, taking the star of a codimension-2 face τ and modding out τ yields a two-dimensional fan. Intuitively, the balancing condition requires that this two-dimensional fan is balanced in the same sense as described above.

The following correspondence between CPWL functions and balanced complex follows from Tropical Geometry (Maclagan and Sturmfels, 2015, Proposition 3.3.10; Maclagan and Sturmfels, 2015, Proposition 3.3.2), as Brandenburg et al. (2025) outlines for the scalar-valued case.

Lemma 18. *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ 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*

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

defines a weight function c_f such that $(\mathcal{C}, c_f)$ is a balanced polyhedral complex. Conversely, for a balanced complex $(\mathcal{C}, c)$, there exists a unique function f, up to addition of a global affine linear function, that is compatible with $\mathcal{C}$ and satisfies $c_f = c$.

B Polyhedral Geometry of ReLU Networks

This appendix collects additional geometric facts about ReLU networks that are used in Section 3. Our goal here is not to redevelop the full framework of Rolnick and Kording (2020), Grigsby et al. (2023) and Grillo and Montúfar (2026), but only to state the specific tools from those works that will be needed for our visibility result in Theorem 2.

B.1 Canonical Polyhedral Complexes and Bent Hyperplanes

We begin by recalling the canonical polyhedral complex associated with a ReLU network as introduced in Grigsby and Lindsey (2022). For a single ReLU layer $\phi_{W,b}(x) = [Wx+b]_+ : \mathbb{R}^d \to \mathbb{R}^m$, the breakpoints are contained in the hyperplane arrangement $\mathcal{H}_{W,b} = \left\{\{x \in \mathbb{R}^d \mid W_i x + b_i = 0\}\right\}_{i=1}^m$. Each sign pattern $\mathbf{s} = (s_1, \dots, s_m) \in \{+, 0, -\}^m$ defines a polyhedron $P_{\mathbf{s}} \coloneqq \bigcap_{i=1}^m H^{s_i}_{W_i, b_i}$, where H^+ and H^- denote the two closed halfspaces of the hyperplane $\{x \mid W_i x + b_i = 0\}$ and H^0 denotes the hyperplane itself. The collection of all such polyhedra forms the *canonical polyhedral complex* $\mathcal{C}_{W,b}$ of the layer $\phi_{W,b}$. On each cell of $\mathcal{C}_{W,b}$, the map $\phi_{W,b}$ is affine linear.

Now let $\theta \in \Theta_{\mathcal{A}}$ for an architecture $\mathcal{A} = (n_0, \dots, n_L, n_{L+1})$. The canonical polyhedral complex of f_θ is constructed iteratively. Set $\mathcal{C}_{\theta,0} \coloneqq \mathbb{R}^{n_0}$. Suppose that $\mathcal{C}_{\theta,\ell-1}$ has already been defined and that all preactivations up to layer ℓ are affine linear on each polyhedron $R \in \mathcal{C}_{\theta,\ell-1}$. For each neuron (j, ℓ), the restriction $z_j^{(\ell,\theta)}|_R$ is an affine function on R.

If $z_j^{(\ell,\theta)}|_R$ is non-constant, its zero set defines a hyperplane in $\mathrm{aff}(R)$, which we denote by $H_R(\ell, j) \coloneqq \{x \in \mathrm{aff}(R) \mid z_j^{(\ell,\theta)}(x) = 0\}$. If $z_j^{(\ell,\theta)}|_R$ is constant, then it introduces no breakpoint on R and does not affect the subdivision. Refining every cell $R \in \mathcal{C}_{\theta,\ell-1}$ by the hyperplanes $H_R(\ell, 1), \dots, H_R(\ell, n_\ell)$ yields the next complex $\mathcal{C}_{\theta,\ell}$. We define the *canonical polyhedral complex* of the network by $\mathcal{C}_\theta \coloneqq \mathcal{C}_{\theta,L}$.

The polyhedra in $\mathcal{C}_\theta$ are indexed by global activation patterns $\mathbf{s} \in \{+, 0, -\}^{n_1} \times \cdots \times \{+, 0, -\}^{n_L}$, and on each such cell the realized map f_θ is affine linear.

For a neuron (j, ℓ), the associated *bent hyperplane* is

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

Thus $B_{\ell,j}(\theta)$ is the locus where the neuron changes its linear behavior. We write $\mathcal{H}_\ell(\theta) \coloneqq \{B_{\ell,j}(\theta) \mid j \in [n_\ell]\}$ for the collection of bent hyperplanes in layer ℓ.

B.2 Generic Parameters

We now record the precise notion of genericity used throughout the paper. This definition is close to those adopted in the works of Phuong and Lampert (2020), Grillo and Montúfar (2026), Masden (2025).

Definition 19. A parameter $\theta \in \Theta_{\mathcal{A}}$ is called *generic* if it satisfies the following two conditions:

1. Supertransversality: every face $\tau \in \mathcal{C}_\theta^{d-k}$ is contained in exactly k bent hyperplanes;
2. Maximal rank of masked products: for every $\ell \in [L+1]$ and every sequence of index sets $S_i \subseteq [n_i]$ for $i \leq \ell - 1$, the matrix product $W^{(\ell)} D_{S_{\ell-1}} W^{(\ell-1)} \cdots D_{S_1} W^{(1)}$ attains the maximum possible rank, namely $\min\Big(d,\ n_\ell,\ \min_{1\leq i\leq \ell-1} |S_i|\Big)$.

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

Lemma 20. *The set $\tilde{\Theta}_{\mathcal{A}}$ is open and dense in $\Theta_{\mathcal{A}}$.*

B.3 Breakpoint Complexes and Visible Facets

Not every facet of the canonical polyhedral complex contributes a genuine breakpoint of the realized function. We therefore distinguish the full canonical subdivision from the visible part of the network.

For a parameter $\theta \in \Theta_{\mathcal{A}}$, let $B(f_\theta) \subseteq \mathbb{R}^{n_0}$ denote the breakpoint set of the realized function. The *breakpoint complex* is the subcomplex $\mathcal{B}_\theta := \{P \in \mathcal{C}_\theta \mid P \subseteq B(f_\theta)\}$. Its support is exactly the breakpoint set: $|\mathcal{B}_\theta| = B(f_\theta)$. Since f_θ is compatible with the canonical polyhedral complex $\mathcal{C}_\theta$, it induces a weight function $c_\theta \colon \mathcal{C}_\theta^{d-1} \to \mathbb{R}^m$ as in Theorem 1. A facet of $\mathcal{C}_\theta$ is visible if and only if the corresponding weight is nonzero, that is, $\mathcal{B}_\theta^{d-1} = \{\sigma \in \mathcal{C}_\theta^{d-1} \mid c_\theta(\sigma) \neq 0\}$. Thus the breakpoint complex can be viewed as the visible part of the canonical polyhedral complex, equipped with the restricted weight function c_θ.

An important observation is that, for generic parameters, the breakpoint complex does not depend on the specific parameter choice but just on the realized function:

Proposition 21 (Grillo and Montúfar, 2026)**.** *If θ, η are generic and satisfy $f_\theta = f_\eta$, then $\mathcal{B}_\theta = \mathcal{B}_\eta$.*

B.4 One-Hidden-Layer Hyperplane Weights

Compared to deep networks, ReLU networks with a single hidden layer have considerably simpler breakpoint geometry: the visible breakpoints lie on a hyperplane arrangement, and the weight function is constant along each visible hyperplane. This structure underlies the hyperplane representation introduced in Section 4.

The following result is taken from Grillo and Montúfar (2026) and is similar to a result of Petzka et al. (2020).

Proposition 22. *Let $\theta = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ be the parameter of a one-hidden-layer network, and let $\sigma \in \mathcal{C}_\theta^{d-1}$ be a facet. Then $c_{f_\theta}(\sigma) = \sum_{i\in I} W^{(2)}_{:,i} \, \|W_i^{(1)}\|$ with $I = \{i \in [n_1] \mid H_i = \mathrm{aff}(\sigma)\}$, where $H_i = \{x \in \mathbb{R}^d \mid W_i^{(1)} x + b_i^{(1)} = 0\}$.*

As an immediate consequence, if $\sigma, \sigma' \in \mathcal{C}_\theta^{d-1}$ are visible facets contained in the same breakpoint hyperplane H_i, then $c_\theta(\sigma) = c_\theta(\sigma')$. Thus, each visible breakpoint hyperplane carries a well-defined vector weight.

B.5 Bending and Visibility Across Layers

The notions introduced in this subsection have been developed systematically in prior work on the polyhedral geometry of ReLU networks (see especially Phuong and Lampert, 2020, Rolnick and Kording, 2020, Grigsby et al., 2023, Grillo and Montúfar, 2026).

We recall here several local geometric criteria from Grillo and Montúfar (2026) that we will use for our analysis of visibility in Appendix B.6.

Definition 23. Let $(\mathcal{C}, c)$ be a weighted polyhedral 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. The ridge τ is called *bending* if at least one adjacent facet is bending at τ, and *non-bending* otherwise.

The next observation states that ridges arising entirely from a single layer are non-bending.

Lemma 24. *If $\tau \in \mathcal{B}_\theta$ lies only on bent hyperplanes from a single layer $\ell \in [L]$, then τ is non-bending.*

Thus any bending ridge must arise from the interaction of different layers. The converse need not hold in general, so we isolate the following property.

Definition 25. We call a parameter θ *honest* if every non-bending ridge of $\mathcal{B}_\theta$ lies only on bent hyperplanes from a single layer.

We also need a visibility condition excluding cancellations of facets.

Definition 26. We call a parameter θ *cancellation-free* if for every facet $\sigma \in \mathcal{C}_\theta^{d-1}$ one has $\sigma \notin \mathcal{B}_\theta \iff$ there exists a layer $k \in [L]$ such that $\mathbf{s}_{k,j}(\sigma) = -$ for all $j \in [n_k]$.

For the visibility theorem, we only need these properties in the generic case.

Lemma 27. *Generic parameters are honest and cancellation-free.*

We next recall the explicit form of the weight function across a visible facet.

Proposition 28. *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. If $S_k \subseteq [n_k]$ denotes the set of strictly active neurons on the relative interior of σ in layer k, then the gradient jump is given by*

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

Finally, the following lemma identifies, at a bending ridge, which adjacent facets come from the earlier layer.

Lemma 29. *Let θ be supertransversal and cancellation-free. Let $\tau \in \mathcal{B}_\theta^{d-2}$ be a bending ridge, and let $H \in \mathcal{H}_k(\theta)$, and $B \in \mathcal{H}_\ell(\theta)$, with $k < \ell$, be the unique bent hyperplanes containing τ. Let $R \in \mathcal{C}_{\theta,k-1}^d$ be the cell with $\operatorname{relint}(\tau) \subseteq \operatorname{relint}(R)$. 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 $e_{\sigma_1/\tau} = -e_{\sigma_2/\tau}$, and these are precisely the facets contained in H. In particular, $H \cap R = \operatorname{aff}(\sigma_1) \cap R$.*
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 on which f_θ is constant, and this facet is the one contained in H. In particular, $H \cap R = \operatorname{aff}(\sigma) \cap R$.*

B.6 Visibility of First-Layer Hyperplanes in Three-Layer Networks

We now prove the visibility result stated as Theorem 2 in the main text. The key point is that bent hyperplanes from the last hidden layer are always visible in the realized function, and that a first-layer hyperplane becomes rigid along the fiber as soon as it is anchored by the second layer.

We begin with a general observation about sums of CPWL functions.

Lemma 30. *Let $f, g \colon \mathbb{R}^d \to \mathbb{R}^m$ be CPWL functions, and let $B_1 \subseteq B(f)$ be a subset of the breakpoint set that is a pure polyhedral complex of codimension 1. Suppose that $B_1 \cap B(g)$ has codimension at least 2. Then*

$$B_1 \subseteq B(f+g).$$

Proof. We first show that $B_1 \setminus B(g) \subseteq B(f+g)$. Let $x \in B_1 \setminus B(g)$. Because g is a CPWL function and $x \notin B(g)$, the function g is affine-linear in a neighborhood of x. Because f is not affine-linear at x (since $x \in B_1 \subseteq B(f)$), the sum $f+g$ is also not affine-linear at x. Hence $x \in B(f+g)$. Thus, $B_1 \setminus B(g)) \subseteq B(f+g)$. Because B_1 is the support of a pure complex of codimension 1

and $B_1 \cap B(g)$ has codimension 2, the intersection $B_1 \cap B(g)$ must be contained in the closure of $B_1 \setminus B(g)$. Since the breakpoint set $B(f+g)$ is closed, it follows that $B_1 \cap B(g) \subseteq B(f+g)$, proving the claim. □

As a consequence, bent hyperplanes from the final hidden layer always remain visible in the realized function.

Lemma 31. *Let $\mathcal{A} = (n_0, n_1, \ldots, n_L, m)$ and let $\theta \in \Theta_{\mathcal{A}}$ be a supertransversal parameter. Then the bent hyperplanes from the last hidden layer are fully contained in the breakpoint complex, that is, $B_{L,j}(\theta) \subseteq B(f_\theta)$ for each $j \in [n_L]$.*

Proof. It is easy to verify that for a vector-valued CPWL function $f\colon \mathbb{R}^d \to \mathbb{R}^m$, the overall breakpoint set is the union of the breakpoint sets of its coordinate functions: $B(f) = \bigcup_{i\in[m]} B(f_i)$.

For each $i \in [m]$, the i-th coordinate of the output is

$$(f_\theta)_i(x) = b_i^{(L+1)} + \sum_{j\in[n_L]} W_{i,j}^{(L+1)} a_j^{(L)}(x).$$

Hence, the breakpoints of $(f_\theta)_i$ depend on the breakpoints of the final activations $a_j^{(L)}$. For every $j \in [n_L]$, the bent hyperplane $B_{L,j}(\theta)$ is exactly contained in $B(a_j^{(L)})$ and is the support of a polyhedral complex pure of codimension 1. Moreover, by supertransversality, the intersection of any two bent hyperplanes has codimension at least 2.

Applying Lemma 30 iteratively over the sum yields that $B_{L,j}(\theta) \subseteq B((f_\theta)_i) \subseteq B(f_\theta)$, proving the claim. □

With these tools in place, we are ready to prove our theorem on rigidity of visible hyperplanes.

Theorem 2. *Let $\mathcal{A} = (d, n_1, n_2, n_3)$ with $d, n_1 > 1$, and let $\theta, \eta \in \Theta_{\mathcal{A}}$ be generic parameters with $f_\theta = f_\eta$. Let $\mathcal{H}_1(\theta)$ denote the set of first-layer hyperplanes of θ. Then:*

1. *If $H \in \mathcal{H}_1(\theta)$ and there exists a neuron $j \in [n_2]$ together with a point $x \in H$ such that $z_j^{(2,\theta)}(x) > 0$, then $H \in \mathcal{H}_1(\eta)$.*
2. *In particular, if this holds for every $H \in \mathcal{H}_1(\theta)$, then $\mathcal{H}_1(\theta) = \mathcal{H}_1(\eta)$ and $\mathcal{C}_\theta = \mathcal{C}_\eta$.*

Proof of Theorem 2. Let $H \in \mathcal{H}_1(\theta)$. By assumption, there exists $x \in H$ such that $z_j^{(2,\theta)}(x) > 0$. We consider two cases based on the behavior of $z_j^{(2,\theta)}$ on H:

Case 1: There also exists $x' \in H$ such that $z_j^{(2,\theta)}(x') < 0$. By continuity, there must be a point on H where $z_j^{(2,\theta)} = 0$, meaning the second-layer bent hyperplane $B_{2,j}(\theta)$ intersects H. Because θ is supertransversal and honest, this intersection contains a bending ridge $\tau \in \mathcal{C}_\theta$. By Lemma 31, since $B_{2,j}(\theta)$ is from the last hidden layer ($L = 2$), this bending ridge is preserved in the breakpoint complex, so $\tau \in \mathcal{B}_\theta$.

Because $f_\theta = f_\eta$ and both parameters are generic, Proposition 21 implies $\mathcal{B}_\theta = \mathcal{B}_\eta$. Thus, τ is also a bending ridge in $\mathcal{B}_\eta$. Applying Lemma 29 to $\mathcal{B}_\eta$ allows us to uniquely identify which of the adjacent facets at τ originate from the earlier layer (layer 1). Since hyperplanes in the first layer are global affine hyperplanes, tracking this facet uniquely identifies the entire hyperplane H. Therefore, $H \in \mathcal{H}_1(\eta)$.

Case 2: $z_j^{(2,\theta)}(y) \geq 0$ for all $y \in H$. Because θ is cancellation-free and $z_j^{(2,\theta)}(x) > 0$ at some point on H, the first-layer hyperplane H is not canceled; it actively contributes to a gradient change in f_θ and hence appears as global affine hyperplane in $B(f_\theta) = B(f_\eta)$.

In the canonical polyhedral complex of a generic parameter, only the first hidden layer produces unbroken, globally affine hyperplanes (as generic second-layer bent hyperplanes must intersect hyperplanes from the first layer for $d, n_1 > 1$ and hence are bent). Therefore, to reproduce this flat hyperplane in $\mathcal{B}_\eta$, it must originate from the first layer of η. Thus, $H \in \mathcal{H}_1(\eta)$.

Finally, if $H \in \mathcal{H}_1(\eta)$ holds for every $H \in \mathcal{H}_1(\theta)$, then we have $\mathcal{H}_1(\theta) \subseteq \mathcal{H}_1(\eta)$. Because θ and η are generic, their first hidden layers have full rank and the same number of neurons n_1. Thus, the number of hyperplanes is exactly the same, which implies $\mathcal{H}_1(\theta) = \mathcal{H}_1(\eta)$ and therefore $\mathcal{C}_{\theta,1} = \mathcal{C}_{\eta,1}$. By Lemma 31, $\mathcal{C}_\theta$ is obtained by refining $\mathcal{C}_{\theta,1}$ such that $\mathcal{B}_\theta$ is a subcomplex. Since $\mathcal{B}_\eta = \mathcal{B}_\theta$, the entire canonical complexes must be identical: $\mathcal{C}_\theta = \mathcal{C}_\eta$. □

C Layerwise Fibers

C.1 Hyperplane Representations

We prove the lemma on the uniqueness of the hyperplane representation (Definition 3).

Lemma 4. *For every one-hidden-layer network $f_\theta \colon \mathbb{R}^d_{\geq 0} \to \mathbb{R}^m$, there exists a hyperplane representation $(\mathcal{H}, A_R, b_R, c)$. Moreover, this representation fully characterizes the function: for any parameter η, we have $f_\eta = f_\theta$ if and only if $(\mathcal{H}, A_R, b_R, c)$ is also a hyperplane representation of f_η.*

Proof. Let $f_\theta \colon \mathbb{R}^d_{\geq 0} \to \mathbb{R}^m$ be a one-hidden-layer network. Consider the hyperplane arrangement $\mathcal{H} = \{H_1, \dots, H_k\}$ consisting of all hyperplanes induced by the nonlinear neurons of the first hidden layer that intersect the domain $\mathbb{R}^d_{\geq 0}$. By construction, the breakpoint set of f_θ is contained in the support of this arrangement, although some of the hyperplanes may carry zero weight and hence may not contribute actual breakpoints of the realized function.

Since f_θ is CPWL, there exists a full-dimensional region $R \in \mathcal{C}_\mathcal{H} \cap \mathbb{R}^d_{\geq 0}$ on which f_θ is affine linear, say $f_\theta(x) = A_R x + b_R$ for all $x \in R$.

By Proposition 22, the weight function c_θ is constant along each hyperplane H_i of the arrangement. Hence for each $i \in [k]$ the quantity $c(i) \coloneqq c_\theta(\sigma)$ is well defined for any facet $\sigma \subseteq H_i$. If the total contribution of the neurons associated with H_i cancels, then $c(i) = 0$; in that case H_i is present in the hyperplane representation but does not appear in the actual breakpoint set. Thus $(\mathcal{H}, A_R, b_R, c)$ is a hyperplane representation of f_θ.

Now let η be another parameter. Suppose that $(\mathcal{H}, A_R, b_R, c)$ is also a hyperplane representation of f_η. Then f_η and f_θ are both compatible with the same arrangement $\mathcal{H}$, they agree with the same affine-linear map on the same reference region R, and they induce the same weight function c on the facets of $\mathcal{C}_\mathcal{H}$. By Theorem 1, the two functions can differ only by a global affine-linear function, and since they agree on the full-dimensional region R, this difference must vanish identically. Hence $f_\eta = f_\theta$.

Conversely, suppose $f_\eta = f_\theta$. Then any hyperplane representation of f_θ is also a hyperplane representation of f_η: the same arrangement $\mathcal{H}$ still covers the breakpoint set, the same affine map (A_R, b_R) is valid on the reference region R, and the induced weight function agrees because the realized functions coincide. Therefore $(\mathcal{H}, A_R, b_R, c)$ is a hyperplane representation of f_η as well. □

C.2 General Fibers

If $(\{H_1, \dots, H_k\}, A_R, b_R, c)$ is a hyperplane representation of $f_\theta \colon \mathbb{R}^d_{\geq 0} \to \mathbb{R}^m$, then the fiber $\mathcal{S}(\theta)$ consists of all parameters that admit this hyperplane representation. To describe this set, we first consider the different possible roles of the hidden neurons.

Definition 32. For $\theta \in \Theta_{(d,n,m)}$, we call a neuron $i \in [n]$ *dead* if $W_i^{(1,\theta)} x + b_i^{(1,\theta)} \leq 0$ for all $x \in \mathbb{R}^d_{\geq 0}$ and *linear* if $W_i^{(1,\theta)} x + b_i^{(1,\theta)} \geq 0$ for all $x \in \mathbb{R}^d_{\geq 0}$, and it is not dead. A neuron is called *nonlinear* if it is neither dead nor linear.

Given a hyperplane representation, each neuron is either nonlinear and associated with one of the hyperplanes, nonlinear and part of a canceling group of neurons, linear, or dead. Accordingly, we consider an assignment map $\phi \colon [n] \to \mathcal{L}$ with $\mathcal{L} = \{1, \dots, k\} \cup \{k+1, \dots, n\} \cup \{\text{lin}, \text{dead}\}$, which partitions the set of neurons into four subsets:

- $N_{\text{vis}} = \phi^{-1}(\{1, \dots, k\})$: neurons assigned to the visible hyperplanes $H_1, \dots, H_k$.

- $N_{\text{can}} = \phi^{-1}(\{k+1, \dots, n\})$: neurons assigned to *canceling groups*. For each $j \in \{k+1, \dots, n\}$, the neurons $\phi^{-1}(j)$ share a hyperplane in $\mathbb{R}^d_{\geq 0}$ and collectively compute an affine function.
- $N_{\text{lin}} = \phi^{-1}(\{\text{lin}\})$: neurons that are linear (always active) on $\mathbb{R}^d_{\geq 0}$.
- $N_{\text{dead}} = \phi^{-1}(\{\text{dead}\})$: neurons that are dead (always inactive) on $\mathbb{R}^d_{\geq 0}$.

Definition 33 (Canonical fiber representation)**.** For a given hyperplane representation $(\{H_1, \dots, H_k\}, A_R, b_R, c)$, let $a_i \in \mathbb{R}^d$ and $t_i \in \mathbb{R}$ be the unique unit vectors and thresholds such that $H_i = \{\langle a_i, x\rangle + t_i = 0\}$ and $R = \{x \in X \mid \langle a_i, x\rangle + t_i \leq 0 \ \forall i \in [k]\}$. Then, for a fixed assignment ϕ and orientation $o \in \{1, -1\}^n$, we define the semi-algebraic set $V_{(\phi,o)}$ as the set of parameters $(W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ satisfying the following conditions:

1. *Normalization*: For all neurons we fix the norm of the rows: $\|W_i^{(1)}\|_2 = 1 \ \forall i \in [n]$.
2. *Breakpoint Structure*: Nonlinear neurons align with hyperplanes according to their assignment and orientation.
   - For visible neurons, the parameters are fixed as:
   $$(W_i^{(1)}, b_i^{(1)}) = o(i)(a_{\phi(i)}, t_{\phi(i)}) \ \ \forall i \in N_{\text{vis}}. \tag{4}$$
   - For canceling neurons, any two in the same group, $j, i \in \phi^{-1}(\ell)$, share the same hyperplane and are oriented according to o as:
   $$(W_i^{(1)}, b_i^{(1)}) = o(i)o(j)(W_j^{(1)}, b_j^{(1)}) \ \ \forall i, j \in \phi^{-1}(\ell), \ \forall \ell \in \{k+1, \dots, n\}. \tag{5}$$
3. *Orthant Constraints*: For linear and dead neurons:
   $$(W_i^{(1)}, b_i^{(1)}) \geq 0 \ \ \forall i \in N_{\text{lin}}, \quad (W_i^{(1)}, b_i^{(1)}) \leq 0 \ \ \forall i \in N_{\text{dead}}. \tag{6}$$
4. *Weight Consistency*: The output weights for each neuron group j add to the corresponding weight:
   $$\sum_{i \in \phi^{-1}(j)} o(i) W_{:,i}^{(2)} = \begin{cases} c(j) & \text{if } j \in \{1, \dots, k\}, \\ 0 & \text{if } j \in \{k+1, \dots, n\}. \end{cases} \tag{7}$$
5. *Affine Base Map*: The neurons active on the reference region R must match (A_R, b_R). A nonlinear neuron $i \in N_{\text{vis}} \cup N_{\text{can}}$ is active on R if and only if $o(i) = -1$:
   $$\sum_{i \in N_{\text{lin}}} W_{:,i}^{(2)} W_i^{(1)} + \sum_{\substack{i \in N_{\text{vis}} \cup N_{\text{can}} \\ o(i) = -1}} W_{:,i}^{(2)} W_i^{(1)} = A_R, \tag{8}$$
   $$\sum_{i \in N_{\text{lin}}} W_{:,i}^{(2)} b_i^{(1)} + \sum_{\substack{i \in N_{\text{vis}} \cup N_{\text{can}} \\ o(i) = -1}} W_{:,i}^{(2)} b_i^{(1)} + b^{(2)} = b_R. \tag{9}$$

We denote by $\mathcal{S}(\theta)/\sim$ the fiber modulo positive rescaling and neuron permutation symmetries.

With these definitions in place, we can describe the fiber of a one-hidden-layer network as follows.

Theorem 34. *The fiber of a one-hidden-layer network $f_\theta \colon \mathbb{R}^d_{\geq 0} \to \mathbb{R}^m$ is given by*

$$\mathcal{S}(\theta)/\sim \ = \bigcup_{\phi, o} V_{(\phi,o)}/\sim,$$

where the union is taken over all possible assignment maps ϕ and orientations o, and $V_{(\phi,o)}$ is the semi-algebraic set given in Definition 33 for the hyperplane representation of f_θ.

Proof of Theorem 34. First we prove $\bigcup_{\phi,o} V_{(\phi,o)} \subseteq \mathcal{S}(\theta)$. Let $\eta \in V_{(\phi,o)}$. By Lemma 4, it suffices to prove that f_η has the same hyperplane representation $(\{H_1, \dots, H_k\}, A_R, b_R, c)$ as θ. This follows by construction; we nonetheless verify the definition. By the breakpoint structure, each nonlinear neuron of η has a hyperplane that either aligns with one of the hyperplanes $H_1, \dots, H_k$ or that cancels

out with other nonlinear neurons. Specifically, if $\sigma \in \mathcal{C}_\eta^{d-1}$ and $\sigma \subseteq H_j$, then, by Proposition 22, for $j \in [k]$, we have that

$$c_\eta(\sigma) = \sum_{i \in \phi^{-1}(j)} [o(i) W^{(2)}_{:,i}] = c(j).$$

For $j > k$, we have that $c_\eta(\sigma) = 0$, so these hyperplanes do not appear in $B(f_\eta)$.

The orthant constraints ensure N_{lin} and N_{dead} introduce no breakpoints in $\text{int}(\mathbb{R}^d_{\geq 0})$, so $B(f_\eta)$ is covered by $\{H_1, \ldots, H_k\}$.

On R, we have $\langle a_j, x\rangle + t_j \leq 0$ for all $j \in [k]$ and hence a nonlinear neuron i has preactivation $o(i)(\langle a_{\phi(i)}, x\rangle + t_{\phi(i)})$, which is ≥ 0 (active) if and only if $o(i) = -1$. Linear neurons are always active. The last equation then ensures the affine map on R matches (A_R, b_R). Altogether, $(\{H_1, \ldots, H_k\}, A_R, b_R, c)$ is a hyperplane representation of f_η.

Conversely, let $\eta = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}) \in \mathcal{S}(\theta)/\sim$. By the scaling symmetry, assume $\|W_i^{(1)}\|_2 = 1$. We define ϕ by classifying neurons:

- Neuron $i \in N_{\text{vis}}$ if $\{x \mid \langle W_i^{(1)}, x\rangle + b_i = 0\} = H_j \in \mathcal{H}$ with $c(j) \neq 0$.
- Neuron $i \in N_{\text{can}}$ if H_j intersects $\text{int}(\mathbb{R}^d_{\geq 0})$ but $c_\eta(\sigma) = 0$ for $\sigma \subseteq \{x \mid \langle W_i^{(1)}, x\rangle + b_i = 0\}$.
- We set $i \in N_{\text{lin}}$ or $i \in N_{\text{dead}}$ if the neuron is linear or dead, respectively.

For $i \in N_{\text{vis}}$, the orientation $o(i)$ is determined by the requirement $W_i^{(1)} = o(i) a_{\phi(i)}$. For $i \in N_{\text{can}}$ one can choose either orientation such that opposite normals have opposite orientation. Because $f_\theta = f_\eta$, it follows that

$$\sum_{i \in \phi^{-1}(j)} o(i) W^{(2)}_{:,i} = \begin{cases} c(j) & \text{if } j \in \{1, \ldots, k\} \\ 0 & \text{if } j \in \{k+1, \ldots, n\} \end{cases}$$

and the affine part on R must match (A_R, b_R). Thus θ satisfies the constraints of $V_{(\phi,o)}$. □

Remark 35. If the domain is $X = \mathbb{R}^d$ instead of $\mathbb{R}^d_{\geq 0}$, then the situation simplifies as there are no inequality constraints. In this case, each component $V_{(\phi,o)}$ becomes an algebraic variety.

C.3 Generic Fibers

For one-hidden-layer networks with generic parameters, the possible functional roles of hidden neurons simplify, and we have:

1. Every nonlinear neuron $i \in N_{\text{vis}}$ corresponds to a unique hyperplane $H_j \in \mathcal{H}$. Thus $k = |N_{\text{vis}}|$ and the assignment map ϕ restricted to N_{vis} is a permutation.
2. There are no canceling neurons. Thus $N_{\text{can}} = \emptyset$.

Next we prove our result describing the generic fiber of a one-hidden-layer network.

Proposition 5. *Let θ be a generic and minimal parameter. Then the generic fiber is given by the disjoint union $\tilde{\mathcal{S}}(\theta)/\sim = \bigcup_{S \subseteq [k]} \left(K_S \cap \tilde{\Theta}_{\mathcal{A}}\right)/\sim$, where the union is taken over orientations $S \subseteq [k]$, and $K_S = \{p_S\} \times V_S$ is the semi-algebraic set defined as follows:*

1. *p_S: The parameters of the k nonlinear neurons are fixed as follows. Let $o \in \{1, -1\}^k$ be the orientation determined by $S = \{i \in [k] \mid o(i) = -1\}$. Then*

$$(W_i^{(1)}, b_i^{(1)}) = o(i)(a_i, t_i), \quad W^{(2)}_{:,i} = c(i) \quad \forall i \in [k]. \tag{2}$$

2. *V_S: The parameters of the linear neurons indexed by $J = [n] \setminus [k]$ and the output bias vector $b^{(2)}$ are subject to the following semi-algebraic constraints, where A_S and b_S are defined in* (1)*:*
   - normalization and positivity*: $\|W_i^{(1)}\|_2 = 1$ and $(W_i^{(1)}, b_i^{(1)}) \geq 0$ for all $i \in J$;*

- linear factorization: $W^{(2)}_{:,J} W^{(1)}_J = A_S$;
- bias alignment: $W^{(2)}_{:,J} b^{(1)}_J + b^{(2)} = b_S$.

Proof of Proposition 5. By minimality, there are no dead neurons, and by genericity, there are no breakpoint cancellations. Since there are exactly k hyperplanes in $\mathcal{H}$, any generic parameter in the fiber must have exactly k nonlinear neurons, one for each hyperplane, and $n-k$ linear neurons.

For each nonlinear neuron $i \in [k]$, there are exactly two normalized parameter configurations $(W^{(1)}_i, b^{(1)}_i)$ that generate the required hyperplane H_i. These two choices correspond to whether the neuron is active or inactive on the reference region R, which is determined by the orientation $o(i) \in \{1, -1\}$. Thus the subsets $S \subseteq [k]$ encode all possible orientation choices for the nonlinear neurons. Once such a subset S is fixed, the output weights $W^{(2)}_{:,i}$ of the nonlinear neurons are uniquely determined by the hyperplane weights $c(i)$.

After fixing the nonlinear neurons, the remaining affine map (A_S, b_S) must be realized by the linear neurons indexed by $J = [n] \backslash [k]$. The semi-algebraic set V_S consists exactly of the possible parameters of these linear neurons and the output bias that realize this affine map subject to the positivity constraints ensuring that the neurons remain linear on the orthant. Hence each set $K_S = \{p_S\} \times V_S$ is a semi-algebraic subset of the full fiber $\mathcal{S}(\theta)$.

Conversely, every generic parameter in the fiber arises in this way: its nonlinear neurons determine a unique subset $S \subseteq [k]$, and its linear neurons must realize the corresponding affine compensation (A_S, b_S). Therefore every generic parameter in the fiber lies in exactly one of the sets K_S.

It follows that $\tilde{\mathcal{S}}(\theta)/\sim = \bigcup_{S \subseteq [k]} \left(K_S \cap \tilde{\Theta}_{\mathcal{A}} \right) / \sim$, as claimed. □

C.4 Dimension of Generic Fibers

The dimensions and non-emptiness of the fiber components K_S in Proposition 5 are fundamentally tied to the ability to decompose the required linear part A_S as a sum of linear neurons. This is captured by the notion of nonnegative column rank.

Definition 36. Let $A \in \mathbb{R}^{m \times d}$ be a real matrix. The *nonnegative column rank* of A, denoted $\mathrm{rank}_{\mathrm{col},+}(A)$, is the smallest integer k such that there exist matrices $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{k \times d}_{\geq 0}$ satisfying $A = UV$.

Lemma 37. *The semi-algebraic set K_S is non-empty if and only if the nonnegative column rank of A_S is at most $n-k$.*

Proof. By definition, if $K_S \neq \emptyset$, then there exist parameters in V_S such that $W^{(2)}_{:,J} W^{(1)}_J = A_S$ with $W^{(1)}_J \geq 0$ and $\|W^{(1)}_i\|_2 = 1$ for all $i \in J$. Any such point gives a nonnegative factorization $A_S = UV$ with $U = W^{(2)}_{:,J} \in \mathbb{R}^{m \times |J|}$ and $V = W^{(1)}_J \in \mathbb{R}^{|J| \times d}_{\geq 0}$, and hence $\mathrm{rank}_{\mathrm{col},+}(A_S) \leq |J|$.

Conversely, if $\mathrm{rank}_{\mathrm{col},+}(A_S) \leq |J|$, take a nonnegative factorization $A_S = BD$. Normalizing the rows of D and compensating by scaling the columns of B preserves the product and satisfies the constraints of V_S, while $b^{(2)}$ can be chosen to meet the affine condition, so $K_S \neq \emptyset$. □

Lemma 38. *Let $\theta \in \Theta_{(d,n,m)}$ be generic and minimal, and let k denote the number of nonlinear neurons. Then $n - k \leq \min\{d, m+1\}$.*

Proof. Let $J = [n] \setminus [k]$. Since θ is generic and minimal, there are no dead neurons, so the neurons in J are exactly the linear neurons.

Their contribution is an affine map $x \mapsto W^{(2)}_{:,J} W^{(1)}_J x + W^{(2)}_{:,J} b^{(1)}_J + b^{(2)}$. Write $A := W^{(2)}_{:,J} W^{(1)}_J$. Because $W^{(1)}_J \geq 0$, this is a nonnegative factorization of A through $|J| = n - k$ hidden units.

By Corollary 42, if $\mathrm{rank}(A) = d$, then $\mathrm{rank}_{\mathrm{col},+}(A) = d$, and otherwise $\mathrm{rank}_{\mathrm{col},+}(A) \leq \mathrm{rank}(A) + 1 \leq d$ by Proposition 40. Hence $\mathrm{rank}_{\mathrm{col},+}(A) \leq d$. Also Proposition 40 gives $\mathrm{rank}_{\mathrm{col},+}(A) \leq m+1$. Therefore $\mathrm{rank}_{\mathrm{col},+}(A) \leq \min\{d, m+1\}$.

If $n-k > \min\{d, m+1\}$, then the same affine map can be realized by strictly fewer than $n-k$ linear neurons. Replacing the linear neurons in θ by such a smaller realization, while keeping the k nonlinear neurons fixed, yields a strict subarchitecture realizing the same function, contradicting minimality. Thus $n-k \leq \min\{d, m+1\}$. □

Proposition 39. *Let $\theta \in \Theta_{(d,n,m)}$ be generic and minimal and let k denote the number of nonlinear neurons. If with $n-k \neq m+1$, then, for every $S \subseteq [k]$, either $\dim(K_S) = (n-k)^2$ or $K_S = \emptyset$.*

Proof. By Lemma 38, we have $n-k \leq \min\{d, m+1\}$. Consider a fixed subset $S \subseteq [k]$ and let (A_S, b_S) denote the affine part of f_θ after flipping the orientation of neurons in S. Define

$$D_S := \{(U, V, u, v) \in \mathbb{R}^{m\times(n-k)} \times \mathbb{R}^{(n-k)\times d} \times \mathbb{R}^m \times \mathbb{R}^{(n-k)} \mid UV = A_S,\ Uv+u = b_S,\ V, v \geq 0\},$$

and denote the projection of D_S onto the matrix components as

$$M_S := \{(U, V) \in \mathbb{R}^{m\times(n-k)} \times \mathbb{R}^{(n-k)\times d} \mid UV = A_S,\ V \geq 0\}.$$

If $M_S = \emptyset$, then also $D_S = \emptyset$, and $K_S = \emptyset$.

Suppose $M_S \neq \emptyset$ and fix a feasible point $(U_0, V_0) \in M_S$. For any

$$X \in \mathrm{GL}_{n-k}^+ \coloneqq \{X \in \mathrm{GL}_{n-k} \mid X \geq 0\},$$

define $(U, V) = (U_0 X^{-1}, X V_0)$. Then

$$UV = U_0 X^{-1} X V_0 = U_0 V_0 = A_S,$$

so the pair (U, V) satisfies the factorization constraint and hence is contained in M_S. By genericity, V_0 has full rank $n-k$, which implies that the map $X \mapsto (U_0 X^{-1}, X V_0)$ is injective. Since GL_{n-k}^+ contains an open set in GL_{n-k}, this shows that $\dim(M_S) \geq (n-k)^2$. Equality follows from the fact that

$$M_S \subseteq \overline{M_S} = \{(U, V) \in \mathbb{R}^{m\times n-k} \times \mathbb{R}^{(n-k)\times d} \mid UV = A_S\},$$

which is parameterized by GL_{n-k} and hence has dimension $(n-k)^2$.

Next, consider the linear constraint $Uv + u = b_S$. For any choice of $v \in \mathbb{R}_{\geq 0}^{(n-k)}$, there exists a unique $u \in \mathbb{R}^m$ satisfying this equation, namely $u = b_S - Uv$. Therefore, the translation degrees of freedom contribute exactly $n-k$ dimensions, corresponding to the entries of v. Passing from D_S to K_S imposes that each row of V has unit Euclidean norm. This removes exactly $n-k$ degrees of freedom from the GL orbit, while leaving the translation degrees of freedom parametrized by v is unaffected.

Combining these contributions, the dimension of K_S is

$$\dim(K_S) = (n-k)^2 + (n-k) - (n-k) = (n-k)^2.$$

This completes the proof. □

Proposition 40. *For any matrix $A \in \mathbb{R}^{m\times d}$, the nonnegative column rank satisfies $\mathrm{rank}_{\mathrm{col},+}(A) \leq \mathrm{rank}(A) + 1$.*

Proof. Let $r = \mathrm{rank}(A)$, and let $C \in \mathbb{R}^{m\times r}$ be a matrix whose columns form a basis of the column space of A, so that $A = CF$ for some $F \in \mathbb{R}^{r\times d}$.

In $\mathbb{R}^r$, consider a minimal complete simplicial fan with $r+1$ rays $v_1, \ldots, v_{r+1} \in \mathbb{R}^r$. Then each column F_j of F can be written as a nonnegative combination of these rays: $F_j = \sum_{i=1}^{r+1} \lambda_{ij} v_i$ with $\lambda_{ij} \geq 0$. If we let $V \in \mathbb{R}^{r\times(r+1)}$ be the matrix with columns $v_1, \ldots, v_{r+1}$ and define $D := (\lambda_{ij}) \in \mathbb{R}_{\geq 0}^{(r+1)\times d}$, then $F = VD$.

Thus, setting $B := CV \in \mathbb{R}^{m\times(r+1)}$, we have $A = BD$. This is a factorization with nonnegative D and inner dimension $r+1$. Therefore, $\mathrm{rank}_{\mathrm{col},+}(A) \leq r+1 = \mathrm{rank}(A)+1$, as claimed. □

Proposition 41. *Let $A \in \mathbb{R}^{m\times d}$ have rank $r \geq 1$. Then $\mathrm{rank}_{\mathrm{col},+}(A) = r$ if and only if the columns of A are contained in a single simplicial cone in $\mathrm{col}(A)$.*

Proof. Let $C \in \mathbb{R}^{m\times r}$ be a matrix whose columns form a basis of $\mathrm{col}(A)$, so that there exists $F \in \mathbb{R}^{r\times d}$ with $A = CF$.

First, note that any r linearly independent vectors in $\mathbb{R}^r$ generate a simplicial cone. Indeed, if $v_1, \dots, v_r \in \mathbb{R}^r$ are linearly independent, then

$$\mathrm{cone}(v_1, \dots, v_r) := \Big\{ \sum_{i=1}^{r} \lambda_i v_i : \lambda_i \geq 0 \Big\}$$

is a simplicial cone in $\mathbb{R}^r$.

If all columns of F lie in a single such cone, then every column of F is a nonnegative combination of these r vectors, and hence $\mathrm{rank}_{\mathrm{col},+}(A) \leq r$.

Conversely, if the columns of F are not contained in a common simplicial cone, then at least $r+1$ vectors are required to generate a cone containing all columns. Therefore, any factorization $A = BD$ with $D \geq 0$ must have at least $r+1$ columns in B, and we conclude that $\mathrm{rank}_{\mathrm{col},+}(A) = r+1$. □

Corollary 42. *Let $A \in \mathbb{R}^{m\times d}$ with $d \leq m$. Then $\mathrm{rank}_{\mathrm{col},+}(A) \leq d$.*

Proof. Let $r = \mathrm{rank}(A) \leq d$. By Proposition 40, we know that $\mathrm{rank}_{\mathrm{col},+}(A) \leq \mathrm{rank}(A) + 1 = r + 1 \leq d + 1$. If $r < d$, then $r + 1 \leq d$, and $\mathrm{rank}_{\mathrm{col},+}(A) \leq d$. If $r = d$, then A has full column rank. In this case, the columns of A are d linearly independent vectors in $\mathbb{R}^m$ and generate a simplicial cone. Hence, by Proposition 41, $\mathrm{rank}_{\mathrm{col},+}(A) = r = d$. Combining these two cases shows $\mathrm{rank}_{\mathrm{col},+}(A) \leq d$, as claimed. □

Proposition 43. *Let $\theta \in \Theta_{(d,n,m)}$ be generic and minimal and let k denote the number of nonlinear neurons. If $n - k = m + 1$, then, for every $S \subseteq [k]$, we have $\dim(K_S) = m^2 + m + d$.*

Proof. Let M_S be as in the proof of Proposition 39. Since θ is minimal, we have $d \geq m + 1$. By genericity, $\mathrm{rank}(A_S) = m$.

We first compute the dimension of $\overline{M_S} = \{(U, V) \in \mathbb{R}^{m\times(m+1)} \times \mathbb{R}^{(m+1)\times d} \mid UV = A_S\}$. Fix $U \in \mathbb{R}^{m\times(m+1)}$ with $\mathrm{rank}(U) = m$. Then $\dim\ker(U) = 1$, so for each column a_j of A_S, the solution set of $Ux = a_j$ is an affine line. Hence $\{V \in \mathbb{R}^{(m+1)\times d} \mid UV = A_S\}$ is an affine space of dimension d. Therefore $\dim(\overline{M_S}) = m(m+1) + d = m^2 + m + d$.

Now let $Z = \{U \in \mathbb{R}^{m\times(m+1)} \mid \exists v \in \mathbb{R}^{m+1}_{>0} : \ker(U) = \mathrm{span}(v)\}$. This is an open subset of $\mathbb{R}^{m\times(m+1)}$, hence $\dim(Z) = m(m+1) = m^2 + m$.

For $U \in Z$, define $M_S(U) := \{V \in \mathbb{R}^{(m+1)\times d} \mid UV = A_S,\ V_{i,j} \geq 0\}$. We claim that $\dim(M_S(U)) = d$. Indeed, since $\mathrm{rank}(U) = m$, the affine space $\{V \in \mathbb{R}^{(m+1)\times d} \mid UV = A_S\}$ has dimension d. Choose $c > 0$ sufficiently large so that $Y_c = \{V \in \mathbb{R}^{(m+1)\times d} \mid UV = A_S,\ |V_{i,j}| \leq c\}$ has nonempty interior relative to this affine space; in particular, $\dim(Y_c) = d$.

Let $v \in \mathbb{R}^{m+1}_{>0}$ satisfy $\ker(U) = \mathrm{span}(v)$ and set $\delta = \min_i v_i > 0$. Define $\varphi_{c,v} \colon [V_1 \quad \cdots \quad V_d] \mapsto \left[V_1 + \frac{c}{\delta}v \quad \cdots \quad V_d + \frac{c}{\delta}v\right]$. Since $Uv = 0$, the map $\varphi_{c,v}$ preserves the equation $UV = A_S$. Moreover, for every entry we have $(V_j)_i + \frac{c}{\delta}v_i \geq -c + \frac{c}{\delta}\delta = 0$, so $\varphi_{c,v}(Y_c) \subseteq M_S(U)$. Since $\varphi_{c,v}$ is an affine isomorphism, it follows that $\dim(M_S(U)) \geq d$. On the other hand, $M_S(U) \subseteq \{V \in \mathbb{R}^{(m+1)\times d} \mid UV = A_S\}$, and the latter has dimension d, hence $\dim(M_S(U)) = d$.

Therefore $M_S \cap (Z \times \mathbb{R}^{(m+1)\times d})$ projects onto Z with all fibers of dimension d. By the fiber dimension theorem for semi-algebraic sets, $\dim(M_S) \geq m^2 + m + d$. Since $M_S \subseteq \overline{M_S}$ and $\dim(\overline{M_S}) = m^2 + m + d$, we conclude that $\dim(M_S) = m^2 + m + d$.

Finally, passing from M_S to K_S is analogous to the proof of Proposition 39: the translation variables contribute $m + 1$ degrees of freedom, while the row-normalization constraints remove exactly $m + 1$ degrees of freedom, and the output bias is then uniquely determined. Hence $\dim(K_S) = \dim(M_S) = m^2 + m + d$. □

With these tools in place, we are ready to prove our theorem on the dimension of the generic fibers.

Theorem 6. *Let θ be generic and minimal. For any given orientation $S \subseteq [k]$, consider the corresponding semi-algebraic component V_S of the generic fiber defined in Proposition 5. Then:*

1. *If $n-k < \min\{d,m\}$, then $\dim(V_S) = (n-k)^2$ if and only if $S = S_\theta$, and $V_S = \emptyset$ otherwise.*
2. *If $n-k = d \leq m$, then $\dim(V_S) = (n-k)^2$ for all $S \subseteq [k]$.*
3. *If $n-k = m < d$, then $\dim(V_S) = (n-k)^2$ if and only if $\operatorname{cone}(A_S) \subseteq \operatorname{col}(A_S)$ is pointed, and $V_S = \emptyset$ otherwise, where A_S is the compensation defined in* (1)*.*
4. *If $n-k = m+1 \leq d$, then $\dim(V_S) = m^2 + m + d$ for all $S \subseteq [k]$.*

Proof of Theorem 6. Let $J = [n] \setminus [k]$. For all $i \in [k]$ we have that

$$c(i)a_i^\top = o_\theta(i) W^{(2,\theta)}_{:,i} W^{(1,\theta)}_i. \tag{10}$$

Recall that on the reference region R we have that

$$f_\eta(x) = A_R x + b_R$$

for all $\eta \in \mathcal{S}(\theta)$. Further, we have that A_R, b_R split into contributions from the linear and non-linear neurons. For θ, let $o_\theta(i)$ be the hyperplane orientation and S_θ the set of nonlinear active neurons on R. Then $A_R = A_{S_\theta} - \sum_{S_\theta} c(i)a_i^\top$. Now choose a different orientation $o \in \{1,-1\}^k$ of the hyperplanes, and let $S = \{i \in [k] \mid o(i) = -1\}$. This needs to be compensated by the contribution of the linear neurons as follows:

$$\begin{aligned} A_R &= A_{S_\theta} - \sum_{S_\theta} c(i)a_i^\top \\ &= A_{S_\theta} - \sum_{S_\theta \setminus S} c(i)a_i^\top - \sum_{S \bigcap S_\theta} c(i)a_i^\top - \sum_{S \setminus S_\theta} c(i)a_i^\top + \sum_{S \setminus S_\theta} c(i)a_i^\top \\ &= A_{S_\theta} - \sum_{S_\theta \setminus S} c(i)a_i^\top + \sum_{S \setminus S_\theta} c(i)a_i^\top - \sum_{S} c(i)a_i^\top. \end{aligned}$$

Thus, for this new set S, the contribution from the linear neurons is required to be

$$A_S = A_R + \sum_S c(i)a_i^\top = A_{S_\theta} - \sum_{S_\theta \setminus S} c(i)a_i^\top + \sum_{S \setminus S_\theta} c(i)a_i^\top.$$

Using Equation (10), we obtain

$$\begin{aligned} A_S &= W^{(2)} \operatorname{diag}(\mathbb{1}_J) W^{(1)} - \sum_{i \in S_\theta \setminus S} o_\theta(i) W^{(2,\theta)}_{:,i} W^{(1,\theta)}_i + \sum_{i \in S \setminus S_\theta} o_\theta(i) W^{(2,\theta)}_{:,i} W^{(1,\theta)}_i \\ &= W^{(2)} \operatorname{diag}(\mathbb{1}_J) W^{(1)} + \sum_{i \in S_\theta \setminus S} W^{(2,\theta)}_{:,i} W^{(1,\theta)}_i + \sum_{i \in S \setminus S_\theta} W^{(2,\theta)}_{:,i} W^{(1,\theta)}_i \\ &= W^{(2)} (\operatorname{diag}(\mathbb{1}_{J \cup (S \triangle S_\theta)}) W^{(1)}. \end{aligned}$$

Essentially, this means we have to add the part from neurons that are in S_θ but not in S and subtract the part from neurons that are in S but not in S_θ from the old linear part. Similar computation for the constant part. This allows us to describe the fiber as follows: we run through all choices of orienting the hyperplanes, each inducing a A_S and then find all factorizations such that the linear neurons multiply to A_S.

By genericity (Definition 19),

$$\operatorname{rank}(A_S) = \min\{d, m, |J \cup (S_\theta \triangle S)|\}.$$

(1) Assume $n-k < \min\{d,m\}$. If $S \neq S_\theta$, then $|J \cup (S_\theta \triangle S)| > n-k$ and $d, m > n-k$, hence $\operatorname{rank}(A_S) > n-k$. Since $\operatorname{rank}_{\operatorname{col},+}(A_S) \geq \operatorname{rank}(A_S)$, Lemma 37 implies $V_S = \emptyset$. If $S = S_\theta$, then $\operatorname{rank}_{\operatorname{col},+}(A_S) = n-k$ and $V_S \neq \emptyset$. The dimension statement follows from Proposition 39.

(2) Assume $n-k=d\le m$. Then $\operatorname{rank}(A_S)=n-k$ for all S. By Corollary 42, $\operatorname{rank}_{\mathrm{col},+}(A_S)=\operatorname{rank}(A_S)=n-k$. Hence $V_S\neq\emptyset$ for all S, and Proposition 39 yields $\dim(V_S)=(n-k)^2$.

(3) Assume $n-k=m<d$. By genericity (Definition 19) we have $\operatorname{rank}(A_S)=m$. By Lemma 37, $V_S\neq\emptyset$ if and only if $\operatorname{rank}_{\mathrm{col},+}(A_S)\le n-k=m$. Let $C=\operatorname{cone}(A_S)\subseteq\operatorname{col}(A_S)$ denote the conic hull of the columns of A_S. By Proposition 41, $\operatorname{rank}_{\mathrm{col},+}(A_S)=m$ if and only if the columns of A_S are contained in a common simplicial cone in $\operatorname{col}(A_S)$.

We show that C is contained in a simplicial cone if and only if it is pointed. Let $C^\circ=\{y\in\mathbb{R}^m:\langle y,x\rangle\ge 0 \text{ for all } x\in C\}$ denote the polar cone of C. The cone C is pointed if and only if C° is full-dimensional.

- If C is pointed, choose m linearly independent vectors $v_1,\dots,v_m\in\operatorname{int}(C^\circ)$. Let $V:=\operatorname{cone}(v_1,\dots,v_m)\subseteq C^\circ$. Then V is a simplicial, full-dimensional cone contained in C°. Take the polar V° of V. This is a simplicial cone and, by polarity and inclusion reversal, we have $C\subseteq V^\circ$.
- Conversely, if C is not pointed, it contains a line and cannot be contained in any simplicial cone.

Combining this with Lemma 37 and Proposition 39, we conclude

$$\dim(V_S)=(n-k)^2 \quad\Longleftrightarrow\quad \operatorname{cone}(A_S)\text{ is pointed,}$$

and $V_S=\emptyset$ otherwise.

(4) Follows by Proposition 43. □

D Fibers from Layer Composition

The Continuous Symmetry

Lemma 8. *Let $\theta=(W^{(\ell)},b^{(\ell)})_{\ell\in[3]}\in\tilde\Theta_{\mathcal{A}}$ be hiding hyperplane j. Then $T_j(\theta)\subseteq\mathcal{S}(\theta)$.*

Proof of Lemma 8. Let $\eta\in T_j(\theta)$. Then there exists $t>\max_{i\in[n_2]}\frac{b_i^{(2)}}{W_{ij}^{(2)}}$ such that $b^{(1,\eta)}=b^{(1)}+te_j$ and $b^{(2,\eta)}=b^{(2)}-t(W_{1j}^{(2)},\dots,W_{n_2j}^{(2)})^\top$, while all other weights and biases coincide with those of θ.

Consider the second-layer preactivation for an arbitrary input $x\in\mathbb{R}^d$:

$$z_i^{(2,\eta)}(x)=\sum_{k=1}^{n_1}W_{ik}^{(2)}a_k^{(1,\eta)}(x)+b_i^{(2,\eta)}=\left(\sum_{k\neq j}W_{ik}^{(2)}a_k^{(1,\theta)}(x)\right)+W_{ij}^{(2)}\left[W_j^{(1)}x+b_j^{(1)}+t\right]_+ +b_i^{(2,\eta)}.$$

We now prove that $a_i^{(2,\eta)}(x)=a_i^{(2,\theta)}(x)$ by showing that the preactivations $z_i^{(2,\eta)}$ and $z_i^{(2,\theta)}$ coincide whenever they are positive. Set $y=W_j^{(1)}x+b_j^{(1)}$. We use the identity

$$z_i^{(2,\eta)}(x)-z_i^{(2,\theta)}(x)=W_{ij}^{(2)}\big([y+t]_+-[y]_+-t\big).$$

We proceed by case distinction over the activation patterns of the first layer.

1. If $y>0$ and $y+t>0$, then $z_i^{(2,\eta)}(x)-z_i^{(2,\theta)}(x)=0$.
2. If $y>0$ and $y+t\le 0$, then $t<0$ and $z_i^{(2,\eta)}(x)\ge z_i^{(2,\theta)}(x)$. Moreover, it holds that
$$z_i^{(2,\eta)}(x)\le W_{ij}^{(2)}[y+t]_+ +b_i^{(2,\eta)}=b_i^{(2,\eta)}=b_i^{(2)}-tW_{ij}^{(2)}\le 0.$$
For the first inequality we used $W_{ik}^{(2)}\le 0$ for all $k\neq j$; for the second equality we used $y+t\le 0$; and for the last inequality we used $b_i^{(2)}\le 0$, $W_{ij}^{(2)}>0$, and $t\le 0$.

3. If $y \le 0$ and $y + t > 0$, then $t > 0$ and $z_i^{(2,\eta)}(x) \le z_i^{(2,\theta)}(x)$. Moreover, it holds that
$$z_i^{(2,\theta)}(x) \le W_{ij}^{(2)}[y]_+ + b_i^{(2)} = b_i^{(2)} \le 0.$$

4. If $y \le 0$ and $y + t \le 0$, then we have that $z_i^{(2,\theta)}(x) \le 0$ as well as $z_i^{(2,\eta)}(x) \le 0$.

In all four cases, $a_i^{(2,\eta)}(x) = a_i^{(2,\theta)}(x)$, and thus $\eta \in \tilde{\mathcal{S}}(\theta)$. Since η was arbitrary in $T_j(\theta)$, it follows that $T_j(\theta) \subseteq \tilde{\mathcal{S}}(\theta)$, as claimed. □

The Discrete Symmetry

The group $\mathcal{G}_j$ is generated by the singleton sets $\{k\}$ for $k \in \{1, \dots, n_1\} \setminus \{j\}$, i.e., $\mathcal{G}_j = \langle M_{\{k\}} : k \in \{1, \dots, n_1\} \setminus \{j\} \rangle$.

Proposition 9. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [3]} \in \Theta_{\mathcal{A}}$ be normalized and hiding hyperplane j. For $M_I \in \mathcal{G}_j$, let $M_I \cdot (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(3)}) = (M_I W^{(1)}, M_I b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(3)})$. Then $\mathcal{G}_j \cdot \theta \subseteq \mathcal{S}(\theta)$.*

Proof of Proposition 9. It suffices to prove the claim for $I = \{i\}$. Let $i \in [n_1] \setminus \{j\}$ and define
$$\eta \coloneqq M_{\{i\}} \cdot \theta = (M_{\{i\}} W^{(1)}, M_{\{i\}} b^{(1)}, W^{(2)}, b^{(2)}, W^{(3)}, b^{(3)}).$$
The first-layer preactivations under η are
$$z_k^{(1,\eta)}(x) = \begin{cases} z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x), & k = j, \\ -z_k^{(1,\theta)}(x), & k = i, \\ z_k^{(1,\theta)}(x), & k \notin \{i, j\}. \end{cases}$$
The second-layer preactivation is $z_1^{(2,\eta)}(x) = \sum_{k=1}^{n_1} W_{1k}^{(2)} a_k^{(1,\eta)}(x) + b_1^{(2)}$.

We now prove that $a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x)$. Fix a region $R \in \mathcal{C}_{1,\theta}$ and let
$$S = \{k \in [n_1] \mid a_k^{(1,\theta)}(x) > 0 \text{ for } x \in R\}$$
denote the the set of first-layer neurons active on R under θ. Let $x \in R$. We distinguish four cases:

Case 1: $i \notin S$, $j \in S$.

Then $z_i^{(1,\theta)}(x) \le 0$ and $z_j^{(1,\theta)}(x) > 0$. Hence
$$a_i^{(1,\eta)}(x) = \left[-z_i^{(1,\theta)}(x)\right]_+ = -z_i^{(1,\theta)}(x),$$
and
$$a_j^{(1,\eta)}(x) = \left[z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x)\right]_+ = z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x).$$
All other activations remain unchanged. Therefore
$$\begin{aligned} z_1^{(2,\eta)}(x) &= W_{1j}^{(2)}\big(z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x)\big) + W_{1i}^{(2)}\big(-z_i^{(1,\theta)}(x)\big) + \sum_{k \notin \{i,j\}} W_{1k}^{(2)} a_k^{(1,\theta)}(x) + b_1^{(2)} \\ &= W_{1j}^{(2)} a_j^{(1,\theta)}(x) + W_{1i}^{(2)} a_i^{(1,\theta)}(x) + \sum_{k \notin \{i,j\}} W_{1k}^{(2)} a_k^{(1,\theta)}(x) + b_1^{(2)} \\ &= z_1^{(2,\theta)}(x), \end{aligned}$$
where we used that the second-hidden-layer neuron is hiding hyperplane j and is normalized, so that $W_{1j}^{(2)} = +1$ and $W_{1i}^{(2)} = -1$, and that $a_i^{(1,\theta)}(x) = 0$. Hence $a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x)$.

Case 2: $i, j \in S$.

Then $z_i^{(1,\theta)}(x) > 0$ and $z_j^{(1,\theta)}(x) > 0$. Thus

$$a_i^{(1,\eta)}(x) = 0, \qquad a_j^{(1,\eta)}(x) = \left[z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x)\right]_+.$$

We compare the second-layer preactivations:

$$z_1^{(2,\theta)}(x) = W_{1j}^{(2)} z_j^{(1,\theta)}(x) + W_{1i}^{(2)} z_i^{(1,\theta)}(x) + \sum_{k\notin\{i,j\}} W_{1k}^{(2)} a_k^{(1,\theta)}(x) + b_1^{(2)},$$

$$z_1^{(2,\eta)}(x) = W_{1j}^{(2)} \left[z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x)\right]_+ + \sum_{k\notin\{i,j\}} W_{1k}^{(2)} a_k^{(1,\theta)}(x) + b_1^{(2)}.$$

If $z_j^{(1,\theta)}(x) \le z_i^{(1,\theta)}(x)$, then $a_j^{(1,\eta)}(x) = 0$ and hence $z_1^{(2,\eta)}(x) \le 0$. Since $W_{1i}^{(2)} < 0$, also $z_1^{(2,\theta)}(x) \le 0$, and thus

$$a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x) = 0.$$

If $z_j^{(1,\theta)}(x) > z_i^{(1,\theta)}(x)$, then

$$z_1^{(2,\eta)}(x) = W_{1j}^{(2)}\big(z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x)\big) + \sum_{k\notin\{i,j\}} W_{1k}^{(2)} a_k^{(1,\theta)}(x) + b_1^{(2)},$$

which is positive if and only if $z_1^{(2,\theta)}(x)$ is positive. In this case, both expressions coincide, and hence

$$a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x).$$

Case 3: $j \notin S$, $i \notin S$**.**

Then $z_j^{(1,\theta)}(x) \le 0$ and $z_i^{(1,\theta)}(x) \le 0$ for all $x \in R$ and $z_j^{(1,\eta)}(x) = z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x)$. Moreover,

$$a_i^{(1,\eta)}(x) = \left[-z_i^{(1,\theta)}(x)\right]_+ = -z_i^{(1,\theta)}(x) \ge 0.$$

All other neurons are unchanged. If $z_j^{(1,\eta)}(x) \le 0$, then since $W_{1k}^{(2)} < 0$ for all $k \ne j$ and $b_1^{(2)} < 0$, we obtain

$$z_1^{(2,\eta)}(x) \le 0 \quad \text{and} \quad z_1^{(2,\theta)}(x) \le 0.$$

Thus

$$a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x) = 0.$$

On the other hand, if $z_j^{(1,\eta)}(x) > 0$, then since $W_{1k}^{(2)} \in \{1, -1\}$, we have that

$$\sum_{k=1}^{n_2} W_{1k}^{(2)} a_k^{(1,\eta)} = \sum_{k\ne i,j} -a_k^{(1,\eta)} + z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x) + z_i^{(1,\theta)}(x) < 0.$$

Since $b^{(2,\eta)} < 0$, this implies that

$$a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x) = 0.$$

Case 4: $j \notin S$, $i \in S$**.**

Then $z_j^{(1,\theta)}(x) \le 0$ and $z_i^{(1,\theta)}(x) > 0$. Hence

$$z_j^{(1,\eta)}(x) = z_j^{(1,\theta)}(x) - z_i^{(1,\theta)}(x) < 0,$$

so $a_j^{(1,\eta)}(x) = 0$, and

$$a_i^{(1,\eta)}(x) = \left[-z_i^{(1,\theta)}(x)\right]_+ = 0.$$

All other neurons are unchanged. Again, since all contributing weights except possibly j are negative and $b_1^{(2)} < 0$, we obtain

$$z_1^{(2,\eta)}(x) \leq 0 \quad \text{and} \quad z_1^{(2,\theta)}(x) \leq 0.$$

Thus

$$a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x) = 0.$$

In all cases, we conclude

$$a_1^{(2,\eta)}(x) = a_1^{(2,\theta)}(x) \quad \forall x \in R.$$

Since R was arbitrary, this holds for all $x \in \mathbb{R}^d$, and thus $\eta \in \tilde{\mathcal{S}}(\theta)$. Since for $I = \{i_1, \ldots, i_q\}$ we have that $M_I \cdot \theta = M_{\{i_1\}} \cdots M_{\{i_1\}} \cdot \theta$, this implies that $M_I \cdot \theta \in \tilde{\mathcal{S}}(\theta)$, proving the claim. □

E Three-Layer Bottleneck Fibers

In a three-layer network, let H_i denote the hyperplane of neuron $(1, i)$ and B_j the bent hyperplane of neuron $(2, j)$, for $i \in [n_1]$ and $j \in [n_2]$. Moreover, for $P \in \mathcal{C}_{\theta,\ell}$, let $S(P) = (S_1(P), \ldots, S_\ell(P))$, where $S_k(P) \subseteq [n_k]$ is the set of active neurons on P. In the remainder of this section, we let $\mathcal{A} = (d, n_1, n_2, m)$ with $n_1 \leq d$. In this case, for generic parameters, all activation patterns are realized on a non-empty polyhedron.

Lemma 44. *Let $\theta \in \tilde{\Theta}_{\mathcal{A}}$ and assume $n_1 \leq d$. Then:*

1. *For every subset $I \subseteq [n_1]$, there exists a unique region $P \in \mathcal{C}_{\theta,1}^d$ such that $S_1(P) = I$.*
2. *For every pair of disjoint subsets $I, Z \subseteq [n_1]$, there exists a non-empty face $\sigma \in \mathcal{C}_{\theta,1}$ such that $\sigma \subseteq \bigcap_{k \in Z} H_k$ and $S_1(\sigma) = I$*

Proof. The statements follow from established results on hyperplane arrangements (see, e.g., Orlik and Terao, 1992, Chapter 2). We offer a proof for completeness. Let $H_k = \{x \in \mathbb{R}^d \mid z_k^{(1,\theta)}(x) = 0\}$ for $k \in [n_1]$. Since θ is generic, the first-layer hyperplanes $H_1, \ldots, H_{n_1}$ are in general position.

We first prove (2). Let $I, Z \subseteq [n_1]$ be disjoint and set $Y \coloneqq \bigcap_{k \in Z} H_k$. Since the arrangement is generic, $\dim Y = d - |Z|$. For every $j \in [n_1] \setminus Z$, the intersection $H_j \cap Y$ is an affine hyperplane in Y, and these hyperplanes are again in general position. Their number is $n_1 - |Z| \leq d - |Z| = \dim Y$. Hence the induced arrangement on Y has all sign patterns. In particular, there exists a unique region σ of the induced arrangement on Y such that $z_k^{(1,\theta)} > 0 \ \ \forall k \in I$, and $z_k^{(1,\theta)} < 0 \ \ \forall k \in [n_1] \setminus (I \cup Z)$. Then σ is a non-empty face of $\mathcal{C}_{\theta,1}$, $\sigma \subseteq \bigcap_{k \in Z} H_k$, and $S_1(\sigma) = I$. This proves (2).

Now we prove (1). Take $Z = \emptyset$ in (2). Then for every $I \subseteq [n_1]$ there exists a non-empty face $\sigma \in \mathcal{C}_{\theta,1}$ with $S_1(\sigma) = I$ and no vanishing first-layer preactivation. Hence σ is full-dimensional, so it is a region $P \in \mathcal{C}_{\theta,1}^d$. This proves existence.

For uniqueness, a region of $\mathcal{C}_{\theta,1}$ is uniquely determined by the signs of the first-layer preactivations and no preactivation can vanish. If $S_1(P) = I$, then $z_k^{(1,\theta)} > 0 \ \ \forall k \in I$, and $z_k^{(1,\theta)} < 0 \ \ \forall k \notin I$. Thus there is at most one such region. This proves (1). □

E.1 Fibers of Non-Hiding Parameters

We show that the first hidden layer is fixed across the fiber when all its hyperplanes are functionally visible. In a bottleneck architecture, linear independence of its weight vectors ensures that any parallelism among facets of the breakpoint complex can be attributed to a unique first-layer neuron.

Lemma 45. *Let $\theta \in \tilde{\Theta}_{\mathcal{A}}$ be generic.*

1. *Let $P \in \mathcal{C}_{\theta,1}^d$ and let $j \in [n_2]$. If $P \cap B_j(\theta) \neq \emptyset$, then the facet $P \cap B_j(\theta)$ is parallel to a first-layer hyperplane $H_i \in \mathcal{H}_1(\theta)$ if and only if $S_1(P) = \{i\}$.*
2. *Let $\sigma_1, \sigma_2 \in \mathcal{C}_\theta^{d-1}$ be second-layer facets. Then σ_1 and σ_2 are parallel if and only if there exist $i \in [n_1]$ and a region $P \in \mathcal{C}_{\theta,1}^d$ such that $S_1(P) = \{i\}$ and both σ_1 and σ_2 are contained in P.*

Proof. On a fixed region $P \in \mathcal{C}^d_{\theta,1}$, the first-layer activation pattern is constant. Hence each second-layer preactivation $z^{(2,\theta)}_j$ restricts to an affine-linear function on P with linear part $\sum_{k\in S_1(P)} W^{(2)}_{jk} W^{(1)}_k$. If $P \cap B_{2,j}(\theta) \neq \emptyset$, then the facet $P \cap B_{2,j}(\theta)$ is defined inside P by the zero set of this affine-linear function, so its normal vector is given by the expression above.

For (1), the facet $P \cap B_{2,j}(\theta)$ is parallel to H_i if and only if its normal vector is parallel to $W^{(1)}_i$. Since θ is generic, the vectors $W^{(1)}_1, \ldots, W^{(1)}_{n_1}$ are linearly independent and all coefficients $W^{(2)}_{jk}$ are nonzero. Therefore $\sum_{k\in S_1(P)} W^{(2)}_{jk} W^{(1)}_k$ is parallel to $W^{(1)}_i$ if and only if exactly one term appears in the sum, namely the one with index i. This is equivalent to $S_1(P) = \{i\}$. This proves (1).

For (2), first suppose that there exist $i \in [n_1]$ and a region $P \in \mathcal{C}^d_{\theta,1}$ with $S_1(P) = \{i\}$ such that both σ_1 and σ_2 are contained in P. Then, by part (1), both facets are parallel to H_i, and hence parallel to each other.

Conversely, suppose that σ_1 and σ_2 are parallel. Since both arise from the second layer, there exist regions $P_1, P_2 \in \mathcal{C}^d_{\theta,1}$ and indices $j_1, j_2 \in [n_2]$ such that $\sigma_r = P_r \cap B_{2,j_r}(\theta)$ for $r = 1, 2$. By part (1), for each $r \in \{1, 2\}$ there exists $i_r \in [n_1]$ such that

$S_1(P_r) = \{i_r\}$. and σ_r is parallel to H_{i_r}. Since σ_1 and σ_2 are parallel, they are parallel to the same first-layer hyperplane, so $i_1 = i_2 =: i$. Thus both P_1 and P_2 have the same sign pattern, namely the one with a single positive entry at index i. Since first-layer regions are uniquely determined by their sign patterns, it follows that $P_1 = P_2$. Hence both facets are contained in the same region $P := P_1 = P_2$, which satisfies the claimed sign condition. □

We now show that the first hidden layer is generally invariant across the fiber, except for hiding parameters. For two weight matrices $W^{(\ell,\theta)}$ and $W^{(\ell,\eta)}$ we write $W^{(\ell,\theta)} \sim W^{(\ell,\eta)}$ if up to relabeling the rows of $W^{(\ell,\theta)}$ are positive scalings of the rows of $W^{(\ell,\eta)}$.

Lemma 46. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ have no dead neurons and be not hiding. If $\eta \in \tilde{\mathcal{S}}(\theta)/\sim$, then $W^{(1)} \sim W^{(1,\eta)}$ and $b^{(1)} \sim b^{(1,\eta)}$.*

Proof. Since θ is not hiding, for every $i \in [n_1]$ there exists $j \in [n_2]$ and $x \in H_{1,i}$ such that $z^{(2,\theta)}_j(x) > 0$. Hence, by Theorem 2, it follows that $\mathcal{H}_1(\theta) = \mathcal{H}_1(\eta)$. It remains to show that the orientation of the hyperplanes is also determined.

Assume that there is a neuron $i \in [n_2]$ such that $b^{(2)}_i > 0$. Then, by genericity, f_θ is constant only on the region where all first-layer neurons are inactive, which determines the orientation of the hyperplanes. Otherwise, there is a neuron $i \in [n_2]$ with $b^{(2)}_i < 0$ and an index $j \in [n_1]$ such that $W^{(2)}_{ij} > 0$. Let P be the region with $S_1(P) = \{j\}$.

Since $b^{(2)}_i < 0$ and $W^{(2)}_{ij} > 0$, the bent hyperplane B_i intersects P. Let σ be a facet of $P \cap B_i(\theta)$. Then σ is parallel to $H_{1,j}$. Since $f_\eta = f_\theta$, the same facet σ occurs in the breakpoint complex of η. By Lemma 45(1), the first-layer region of η containing σ must again be the region in which only neuron j is active. Therefore the orientation of all first-layer hyperplanes is fixed and this completes the proof. □

This rigidity reduces the analysis of the three-layer fiber to the one-hidden-layer case in Section 4.

Corollary 47. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ satisfy the assumptions of Lemma 46. Then the generic fiber decomposes as*

$$\tilde{\mathcal{S}}(\theta)/\sim \; = (\{(W^{(1)}, b^{(1)})\} \times \tilde{\mathcal{S}}(\theta'))/\sim,$$

where $\tilde{\mathcal{S}}(\theta')$ is the generic fiber of the one hidden layer network $g_{\theta'} : \mathbb{R}^{n_1}_{\geq 0} \to \mathbb{R}^m$ given by $g_{\theta'}(y) = W^{(3)}[W^{(2)}y + b^{(2)}]_+ + b^{(3)}$, as characterized in Proposition 5.

Proof. By Lemma 46, the first-layer parameters are fixed across $\tilde{\mathcal{S}}(\theta)$ up to trivial symmetry. The condition $f_\eta = f_\theta$ therefore requires the remaining layers to agree on the image of the first layer.

Since $n_1 \leq d$ and $W^{(1)}$ has full rank, the map $x \mapsto [W^{(1)}x + b^{(1)}]_+$ is surjective onto the non-negative orthant $\mathbb{R}^{n_1}_{\geq 0}$. Hence, the fiber is determined by functional equality on $\mathbb{R}^{n_1}_{\geq 0}$, reducing the problem to the case analyzed in Section 4. □

E.2 Fibers of Hiding Parameters

We now characterize the fiber when a hyperplane is hidden. Even in this case, the weights $W^{(1)}$ remain invariant if the second layer is sufficiently wide ($n_2 \geq 2$), as multiple second-layer neurons provide enough "linear evidence" of the hidden hyperplane's direction.

Lemma 48. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [3]} \in \tilde{\Theta}_{\mathcal{A}}$ be generic and hiding hyperplane j. Let $T_j(\theta)$ be given as in Equation* (3)*. Then $\{\eta \in \tilde{\mathcal{S}}(\theta) \mid W^{(1,\eta)} = W^{(1)}\} = T_j(\theta) \cap \tilde{\Theta}_{\mathcal{A}}$.*

Proof. We first show $T_j(\theta) \cap \tilde{\Theta}_{\mathcal{A}} \subseteq \{\eta \in \tilde{\mathcal{S}}(\theta) \mid W^{(1,\eta)} = W^{(1)}\}$. Let $\eta \in T_j(\theta) \cap \tilde{\Theta}_{\mathcal{A}}$. By Lemma 8, we have $\eta \in \mathcal{S}(\theta)$. Since moreover $\eta \in \tilde{\Theta}_{\mathcal{A}}$, it follows that $\eta \in \tilde{\mathcal{S}}(\theta)$. By construction, elements of $T_j(\theta)$ preserve the first-layer weights, so $W^{(1,\eta)} = W^{(1)}$.

For the converse inclusion, let $\eta \in \tilde{\mathcal{S}}(\theta) \cap \{\eta \in \tilde{\Theta}_{\mathcal{A}} \mid W^{(1,\eta)} = W^{(1)}\}$. Since the first-layer weights are fixed, the second-layer weights $W^{(2)}$ are also fixed across $\tilde{\mathcal{S}}(\theta)$. Indeed, on any region where all first-layer neurons are active (which exists by genericity and the assumption $n_1 \leq d$), the preactivations $z_i^{(\eta,2)}$ are affine functions with linear part $W_i^{(2,\eta)}W^{(1)}$, which uniquely determines $W_i^{(2,\eta)}$ since $W^{(1)}$ has full rank. Now, normalizing so that $\|W_i^{(2)}W^{(1)}\| = 1$, Proposition 28 implies that the output weights $W_{:i}^{(3)}$ are also uniquely determined. Hence $W^{(3)}$ is also fixed across $\tilde{\mathcal{S}}(\theta)$.

By Theorem 2, all first-layer hyperplanes, except possibly the one corresponding to neuron j, are fixed across the fiber. Thus there exists $t \in \mathbb{R}$ such that $b^{(1,\eta)} = b^{(1)} + te_j$.

Since θ is hiding hyperplane j, for every $i \in [n_2]$ we have $W_{ij}^{(2)} > 0$, $W_{ik}^{(2)} < 0$ for $k \neq j$, and $b_i^{(2)} < 0$. Let P be the unique region of $\mathcal{C}_{\theta,1}$ with $S_1(P) = \{j\}$ Then, each second-layer bent hyperplane intersects P, and by Lemma 45, the facets

$$\sigma_i = P \cap B_i(\theta) = \{x \in \mathbb{R}^d \mid W_{ij}^{(2)}(W_j^{(1)}x + b_j^{(1)}) + b_i^{(2)} = 0\}$$

are all parallel to H_j. By the same lemma, these facets σ_i must be contained in some region $R \in \mathcal{C}_{\eta,1}$ on which only neuron j is active. Moreover, comparing with the corresponding expression for θ shows that necessarily $b_i^{(2,\eta)} = b_i^{(2)} - tW_{ij}^{(2)}$ for all $i \in [n_2]$.

Since $W_{ij}^{(2)} > 0$ and $b_i^{(2)} < 0$, the region R exists only if $b_i^{(2,\eta)} < 0$ for all i, i.e., $b_i^{(2)} - tW_{ij}^{(2)} < 0$. This is equivalent to $t > \max_{i \in [n_2]} \frac{b_i^{(2)}}{W_{ij}^{(2)}}$. By the same argument as in the forward inclusion, this implies that $a_i^{(2,\eta)}(x) = a_i^{(2,\theta)}(x)$ for all i, hence $\eta \in T_j(\theta) \cap \tilde{\Theta}_{\mathcal{A}}$. □

Lemma 49. *Let θ be generic and suppose that neuron $i \in [n_2]$ hides hyperplane $j \in [n_1]$. Then for every face $\sigma \in \mathcal{C}_{\theta,1}$ we have that $B_i \cap \sigma \neq \emptyset$ if and only if $j \in S_1(\sigma)$.*

Proof. Let $\sigma \in \mathcal{C}_{\theta,1}$ be a face. On σ, the active set $S_1(\sigma)$ is constant, and the second-layer preactivation of neuron i restricts to $z_i^{(2,\theta)}(x) = \sum_{k \in S_1(\sigma)} W_{ik}^{(2)} z_k^{(1,\theta)}(x) + b_i^{(2)}$ for all $x \in \sigma$.

Assume first that $j \notin S_1(\sigma)$. Since neuron i hides hyperplane j, we have $W_{ik}^{(2)} \leq 0$ for all $k \neq j$ and $b_i^{(2)} < 0$ by genericity (otherwise the second-layer bent hyperplane corresponding to neuron i would intersect the intersection of the first-layer hyperplanes violating supertransversality). Moreover, $z_k^{(1,\theta)}(x) = a_k^{(1,\theta)}(x) > 0$ for all $k \in S_1(\sigma)$ and all $x \in \sigma$. Hence $z_i^{(2,\theta)}(x) = \sum_{k \in S_1(\sigma)} W_{ik}^{(2)} a_k^{(1,\theta)}(x) + b_i^{(2)} < 0$ for all $x \in \sigma$. Therefore $z_i^{(2,\theta)}$ has no zero on σ, and thus $B_i \cap \sigma = \emptyset$.

Now assume that $j \in S_1(\sigma)$. By Lemma 44, there exists a non-empty face $\tau \in \mathcal{C}_{\theta,1}$ such that $\tau \subseteq \sigma \cap \left(\bigcap_{k \in S_1(\sigma)\setminus\{j\}} H_k\right)$ and $S_1(\tau) = \{j\}$. For all $x \in \tau$ we therefore have $z_i^{(2,\theta)}(x) = W_{ij}^{(2)} z_j^{(1,\theta)}(x) + b_i^{(2)}$. At $\tau \cap H_j$, this value equals $b_i^{(2)} < 0$.

On the other hand, since τ is unbounded in the direction where only neuron j remains active, the function $z_j^{(1,\theta)}$ is unbounded above on τ. Hence there exists $x_+ \in \tau$ such that $W_{ij}^{(2)} z_j^{(1,\theta)}(x_+) > -b_i^{(2)}$, and therefore $z_i^{(2,\theta)}(x_+) > 0$. By continuity, $z_i^{(2,\theta)}$ must vanish at some point of τ, and since $\tau \subseteq \sigma$, it follows that $B_i \cap \sigma \neq \emptyset$, proving the claim. □

Proposition 11. *Let $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ be generic and hiding hyperplane j, and assume that $n_2 \geq 2$. Then $\tilde{\mathcal{S}}(\theta)/\sim = (T_j(\theta) \cap \tilde{\Theta}_{\mathcal{A}})/\sim$.*

Proof of Proposition 11. By Lemma 48, it suffices to show that $\eta \in \tilde{\mathcal{S}}(\theta)/\sim$ implies $W^{(1,\eta)} \sim W^{(1)}$. By Theorem 2 and Lemma 49, all first-layer hyperplanes of η coincide with those of θ except possibly for the j-th hyperplane. Thus $\mathcal{H}_1(\eta) = (\mathcal{H}_1(\theta) \setminus \{H_j\}) \cup \{\widehat{H}_j\}$ for some hyperplane $\widehat{H}_j$.

By Lemma 44, let $P \in \mathcal{C}_{\theta,1}$ be the region where only neuron j is active in the first layer. It follows from Lemma 45 that for any second-layer neuron $i \in [n_2]$ the facet

$$\sigma_i = P \cap B_i(\theta) = \{x \in P \mid W_{ij}^{(2)}(W_j^{(1)}x + b_j^{(1)}) + b_i^{(2)} = 0\}$$

is parallel to $H_j = \{x \mid W_j^{(1)}x + b_j^{(1)} = 0\}$. Since $n_2 \geq 2$ and by Lemma 49 all bent hyperplanes intersect P, there are at least two such parallel facets inside P. As these facets must be facets of the breakpoint complex for each parameter realizing the same function, it follows again by Lemma 45 that $\widehat{H}_j$ is parallel to H_j. Moreover, by the same lemma, only neuron j can be active on the region containing the two parallel facets, which implies that the orientation of the hyperplanes is also determined. Thus, the first-layer weight matrix $W^{(1)}$ is fixed (up to trivial symmetries) across $\tilde{\mathcal{S}}(\theta)$. The claim now follows from Lemma 48. □

Proposition 12. *Let $\theta = (W^{(\ell,\theta)}, b^{(\ell,\theta)})_{\ell\in[3]} \in \tilde{\Theta}_{\mathcal{A}}$ be generic and normalized, assume that $n_2 = 1$, and suppose that θ hides hyperplane j. Then $\tilde{\mathcal{S}}(\theta)/\sim = \left(\bigcup_{I\subseteq[n_1]\setminus\{j\}} \left(T_j(M_I \cdot \theta) \cap \tilde{\Theta}_{\mathcal{A}}\right)\right)/\sim$.*

Proof of Proposition 12. We first show $\left(\bigcup_{I\subseteq[n_1]\setminus\{j\}} \left(T_j(M_I \cdot \theta) \cap \tilde{\Theta}_{\mathcal{A}}\right)\right)/\sim \;\subseteq\; \tilde{\mathcal{S}}(\theta)/\sim$.

Let η belong to the left-hand side. Then there exist $I \subseteq [n_1]\setminus\{j\}$ and $\widetilde{\eta} \in T_j(M_I \cdot \theta) \cap \tilde{\Theta}_{\mathcal{A}}$ such that $\eta \sim \widetilde{\eta}$. By Proposition 9, we have $M_I \cdot \theta \in \mathcal{S}(\theta)$, and by Lemma 8, we have $T_j(M_I \cdot \theta) \in \mathcal{S}(M_I \cdot \theta)$. Hence $\widetilde{\eta} \in \mathcal{S}(\theta)$. Moreover, since $\widetilde{\eta} \in \tilde{\Theta}_{\mathcal{A}}$, it follows that $\widetilde{\eta} \in \tilde{\mathcal{S}}(\theta)$. Therefore $\eta \in \tilde{\mathcal{S}}(\theta)/\sim$.

For the **reverse inclusion**, let $\eta \in \tilde{\mathcal{S}}(\theta)/\sim$. We will show that $\eta \in T_j(M_I \cdot \theta)$ for some $I \subseteq [n_1]\setminus\{j\}$.

η also hides hyperplane j: Since θ hides hyperplane j and $f_\eta = f_\theta$, Theorem 2 and Lemma 49 imply that $H_k(\eta) = H_k(\theta)$ for all $k \neq j$. Thus, only the j-th hyperplane (the hidden hyperplane) may differ. Because $f_\eta = f_\theta$, the breakpoint sets agree: $B(f_\eta) = B(f_\theta)$. Since $H_j(\theta)$ is hidden in θ, it does not occur in $B(f_\theta)$. Since η is generic, the corresponding first-layer hyperplane of η is also hidden, which implies $W_{1j}^{(2,\eta)} > 0$, $W_{1k}^{(2,\eta)} < 0$ for all $k \neq j$, and $b_1^{(2,\eta)} < 0$, by Lemma 49. Replacing η by an equivalent representative if necessary, we may assume that η is normalized, meaning that the second layer weights are either 1 or -1. Hence, $W_{1j}^{(2,\eta)} = 1$ and $W_{1k}^{(2,\eta)} = -1$ for all $k \neq j$.

Possible choices for $W_j^{(1,\eta)}$: Because $f_\eta = f_\theta$, the visible second-layer bent hyperplane agrees: $B_1(\eta) = B_1(\theta)$. For any first-layer region $Q \in \mathcal{C}_{\theta,1}$ with active set $S \subseteq [n_1]$, the linear part of $z_1^{(2,\theta)}$ on Q is $\sum_{k\in S} W_{1k}^{(2,\theta)} W_k^{(1,\theta)}$. Since θ is normalized and hides hyperplane j, we have $W_{1j}^{(2,\theta)} = 1$ and $W_{1k}^{(2,\theta)} = -1$ for $k \neq j$. Hence every visible facet of $B_1(\theta)$ has normal vector of the form $W_j^{(1,\theta)} - \sum_{k\in I} W_k^{(1,\theta)}$, for some $I \subseteq [n_1] \setminus \{j\}$.

By Lemma 44, there exists a unique region $R \in \mathcal{C}^d_{\eta,1}$ on which only neuron j is active under η. Since η is normalized, the linear part of $z_1^{(2,\eta)}$ on R is $W_{1j}^{(2,\eta)} W_j^{(1,\eta)} = W_j^{(1,\eta)}$. Let $\sigma = R \cap B_1(\eta)$, which is not empty by Lemma 49 because η hides hyperplane j. Because $B_1(\eta) = B_1(\theta)$, the same facet σ is a visible facet of $B_1(\theta)$. Hence there exists $I \subseteq [n_1] \setminus \{j\}$ such that $W_j^{(1,\eta)} = W_j^{(1,\theta)} - \sum_{k \in I} W_k^{(1,\theta)}$.

Compensate the choice of weights: Let $Q_{\mathrm{all}} \in \mathcal{C}_{\theta,1}$ be the unique region on which all first-layer neurons are active under θ, which again exists by Lemma 44. On this region, the linear part of $z_1^{(2,\theta)}$ is $W_j^{(1,\theta)} - \sum_{k \in [n_1] \setminus \{j\}} W_k^{(1,\theta)}$. Let $\sigma' = Q_{\mathrm{all}} \cap B_1(\theta)$. Since $f_\eta = f_\theta$, the same facet σ' occurs in the breakpoint complex of η. Let $Q_\eta \in \mathcal{C}_{\eta,1}$ be the first-layer region containing σ', and let $S_1(Q_\eta) \subseteq [n_1]$ be the set of active neurons on Q_η. For each $k \neq j$, the hyperplanes $H_k(\eta) = H_k(\theta)$ coincide, so there exists $\varepsilon_k \in \{\pm 1\}$ such that $W_k^{(1,\eta)} = \varepsilon_k W_k^{(1,\theta)}$ for all $k \neq j$.

Since η is normalized, the linear part of $z_1^{(2,\eta)}$ on Q_η is

$$\mathbf{1}_{\{j \in S_1(Q_\eta)\}} W_j^{(1,\eta)} - \sum_{k \in S_1(Q_\eta) \setminus \{j\}} W_k^{(1,\eta)}, \tag{11}$$

where

$$\mathbf{1}_{\{j \in S_1(Q_\eta)\}} = \begin{cases} 1 & j \in S_1(Q_\eta), \\ 0 & j \notin S_1(Q_\eta). \end{cases}$$

Substituting $W_j^{(1,\eta)} = W_j^{(1,\theta)} - \sum_{k \in I} W_k^{(1,\theta)}$ and $W_k^{(1,\eta)} = \varepsilon_k W_k^{(1,\theta)}$ for $k \neq j$, we obtain that Expression 11 is equal to

$$\mathbf{1}_{\{j \in S_1(Q_\eta)\}} \left(W_j^{(1,\theta)} - \sum_{k \in I} W_k^{(1,\theta)} \right) - \sum_{k \in S_1(Q_\eta) \setminus \{j\}} \varepsilon_k W_k^{(1,\theta)}.$$

Since σ' is also a facet of $Q_{\mathrm{all}} \cap B_1(\theta)$, this must also equal $W_j^{(1,\theta)} - \sum_{k \in [n_1] \setminus \{j\}} W_k^{(1,\theta)}$. Because the vectors $W_1^{(1,\theta)}, \ldots, W_{n_1}^{(1,\theta)}$ are linearly independent, it follows that $j \in S_1(Q_\eta)$ and $S_1(Q_\eta) = [n_1] \setminus I$. Indeed, if $k \in I$, then the coefficient of $W_k^{(1,\theta)}$ is already -1 from the first sum, so $k \notin S_1(Q_\eta)$. If $k \notin I$, then k must belong to $S_1(Q_\eta)$ in order to contribute the coefficient -1 on the right-hand side.

Now Q_{all} is the unique region in which every fixed neuron $k \neq j$ is active under θ. Since the hyperplanes $H_k(\eta) = H_k(\theta)$ are fixed and $S_1(Q_\eta) = [n_1] \setminus I$, it follows that the orientation of neuron k is preserved for $k \notin I$ and reversed for $k \in I$. Therefore $W^{(1,\eta)} = M_I W^{(1,\theta)}$.

Since both η and $M_I \cdot \theta$ are normalized and hide hyperplane j, their second-layer weights coincide. Since the realized functions agree, by Proposition 28, the output weights coincide as well. Therefore Lemma 48 applied to $M_I \cdot \theta$ implies that η differs from $M_I \cdot \theta$ only by a translation of the hidden hyperplane. Hence $\eta \in T_j(M_I \cdot \theta) \cap \tilde{\Theta}_{\mathcal{A}}$, and therefore $\eta \in \left(\bigcup_{I \subseteq [n_1] \setminus \{j\}} \left(T_j(M_I \cdot \theta) \cap \tilde{\Theta}_{\mathcal{A}} \right) \right) / \sim$. This proves the reverse inclusion. □

E.3 Characterization of Identifiability

Proposition 14. *There exists a polynomial-time algorithm that, given a bottleneck architecture $\mathcal{A} = (d, n_1, n_2, m)$ with $n_1 \leq d$ and two generic parameters $\theta, \eta \in \tilde{\Theta}_{\mathcal{A}}$, decides whether $f_\theta = f_\eta$.*

Proof of Proposition 14. We describe a decision procedure and verify that each step can be carried out in polynomial time.

First inspect the sign pattern of the rows of $[W^{(2)}, b^{(2)}]$ in order to determine whether θ and η are hiding or non-hiding. Moreover, dead neurons can be removed. By the bottleneck and genericity assumption, the image of the first hidden layer is always the entire nonnegative orthant and hence a neuron is dead if and only if all its incoming weights and its bias is nonpositive.

Case 1: non-hiding. By Lemma 46, if $f_\theta = f_\eta$, then the first-layer parameters of θ and η must agree up to permutation and positive rescaling of rows. Thus we first compare the first-layer hyperplanes

of θ and η, equivalently, we check whether the rows of $[W^{(1)}, b^{(1)}]$ match up to permutation and positive scaling. If not, then $f_\theta \neq f_\eta$.

Assume now that the first layers agree in this sense. By Corollary 47, equality of f_θ and f_η is equivalent to equality of the induced one-hidden-layer networks on $\mathbb{R}^{n_1}_{\geq 0}$. By Lemma 4, this is equivalent to equality of the corresponding weighted hyperplane arrangements together with equality of the affine map on one base region. These quantities are explicitly computable from the parameters using sign comparisons and linear-algebraic operations, and can therefore be checked in polynomial time.

Case 2: hiding with $n_2 \geq 2$. By Proposition 11, if $f_\theta = f_\eta$, then both parameters hide the same first-layer hyperplane H_j, all remaining first-layer hyperplanes agree up to permutation and positive rescaling, and η differs from θ only by a translation of the hidden hyperplane. Thus we first identify the hidden index j from the sign pattern of $[W^{(2)}, b^{(2)}]$ and check that the same index occurs for η. We then compare all non-hidden first-layer hyperplanes. If these checks fail, then $f_\theta \neq f_\eta$.

If they succeed, then by Lemma 48 equivalence is equivalent to the existence of a scalar t such that the first- and second-layer biases differ exactly as in Equation (3). This is a system of linear equalities and inequalities in one unknown t, and can therefore be checked in polynomial time.

Case 3: hiding with $n_2 = 1$. Again we first identify the hidden index j and compare all non-hidden first-layer hyperplanes. Their relative orientations determine the corresponding sign flips. By Proposition 12, once the non-hidden hyperplanes are matched, the first-layer parameter of the hidden neuron is uniquely determined by the visible second-layer bent hyperplane. Equivalently, the subset $I \subseteq [n_1] \setminus \{j\}$ in the transformation M_I is recovered from the orientations of the matched non-hidden hyperplanes. Thus no search over subsets is required.

Having recovered I, equivalence is equivalent to checking whether η differs from $M_I \cdot \theta$ by a translation of the hidden hyperplane as in Equation (3). As in Case 2, this reduces to checking a system of linear equalities and inequalities in one scalar t, and is therefore polynomial-time decidable.

These three cases exhaust all generic parameters in the bottleneck setting. Hence the procedure decides whether $f_\theta = f_\eta$ in polynomial time. □

Corollary 15. *Let $\mathcal{A} = (d, n_1, n_2, m)$ with $d \geq n_1 \geq 2$, let $\mathcal{I} \subseteq \Theta_\mathcal{A}$ be the set of identifiable parameters, and $\mathbb{B}_r$ the radius-r ball in $\Theta_\mathcal{A}$ for an arbitrary $r > 0$. Then*

$$1 - n_2\, 2^{-n_1} - n_1\, 2^{-n_2 n_1} \;\leq\; \frac{\mathrm{vol}(\mathcal{I} \cap \mathbb{B}_r)}{\mathrm{vol}(\Theta_\mathcal{A} \cap \mathbb{B}_r)} \;\leq\; (1 - 2^{-(n_1+1)})^{n_2}.$$

Proof of Corollary 15. By Theorem 13, non-identifiability of a generic parameter is completely determined by the sign pattern of the rows of $[W^{(2)}, b^{(2)}]$. Since generic parameters form a full-measure subset of $\Theta_\mathcal{A}$ and all orthants have the same relative volume, it suffices to compute probabilities under the uniform distribution on sign patterns.

A necessary condition for identifiability is that no row of $[W^{(2)}, b^{(2)}]$ only has negative entries. For a fixed row, this happens with probability $2^{-(n_1+1)}$, and for all rows these events are independent. Hence,

$$\mathbb{P}(\text{no row is all negative}) = (1 - 2^{-(n_1+1)})^{n_2}.$$

Since identifiability implies this condition, we obtain

$$\frac{\mathrm{vol}(\mathcal{I} \cap \mathbb{B}_r))}{\mathrm{vol}(\Theta_\mathcal{A} \cap \mathbb{B}_r))} \leq (1 - 2^{-(n_1+1)})^{n_2}.$$

By Theorem 13, a parameter is non-identifiable if at least one of the statements 1,2 or 3 holds. We bound the probability of each event.

Event 1. For each row, the probability that all entries are negative is $2^{-(n_1+1)}$. By the union bound over the n_2 rows,

$$\mathbb{P}(\text{Event 1}) \leq n_2\, 2^{-(n_1+1)}.$$

Event 2. For each row, the probability that all entries are positive is again $2^{-(n_1+1)}$, implying

$$\mathbb{P}(\text{Event 2}) \leq n_2\, 2^{-(n_1+1)}.$$

Event 3. Fix $j \in [n_1]$. For a single row, the probability of being positive in column j and negative elsewhere is $2^{-(n_1+1)}$. By independence across rows, the probability that all n_2 rows satisfy this pattern is $2^{-n_1 n_2}$. Taking a union bound over $j \in [n_1]$ yields

$$\mathbb{P}(\text{Event 3}) \leq n_1\, 2^{-n_1 n_2}.$$

By the union bound,

$$\mathbb{P}(\text{non-identifiable}) \leq \sum_{i=1}^{3} \mathbb{P}(\text{Event i}),$$

and therefore

$$\frac{\operatorname{vol}(\mathcal{I} \cap \mathbb{B}_r))}{\operatorname{vol}(\Theta_{\mathcal{A}} \cap \mathbb{B}_r))} = 1 - \mathbb{P}(\text{non-identifiable}) \geq 1 - 2n_2\, 2^{-(n_1+1)} - n_1\, 2^{-n_2 n_1} = 1 - n_2\, 2^{-n_1} - n_1\, 2^{-n_2 n_1}.$$

Combining the upper and lower bounds proves the claim. □

F Implications for Deeper Networks

F.1 Fibers Are Not Localizable

Proposition 16. *Let $\mathcal{A} = (n_0, n_1, n_2, n_3, m)$ with $n_0 \geq n_1 \geq n_2$. Then there exists $\theta = (W^{(\ell)}, b^{(\ell)})_{\ell \in [4]}$ such that, for the layer maps $f_\ell \colon \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$ defined by $x \mapsto \left[W^{(\ell)} x + b^{(\ell)}\right]_+$, the compositions $f_2 \circ f_1$ and $f_3 \circ f_2$ are identifiable, while $f_3 \circ f_2 \circ f_1$ is not.*

Proof of Proposition 16. The idea is to construct parameters that satisfy the sign condition for identifiability from Theorem 13, while ensuring that the image of $f_2 \circ f_1$ is bounded so that the neurons of f_3 act linearly on this image.

Let $W^{(1)} = I_d$, $b^{(1)} = 0$, $W^{(2)}_{ij} < 0$, $b^{(2)} \geq 0$, and $W^{(3)}_{ij} < 0$, $b^{(3)} = tv$ for some positive vector v and positive scalar t. All these layers and also the fourth layer are chosen generically in the sense of Definition 19.

By construction, the rows of $[W^{(2)}, b^{(2)}]$ and $[W^{(3)}, b^{(3)}]$ have mixed signs, no row is strictly negative nor strictly positive, and no column satisfies the hiding condition of Theorem 13. Hence $f_2 \circ f_1$ and $f_3 \circ f_2$ are identifiable. However, the image of $f_2 \circ f_1$ is bounded.

Hence, choosing $t > 0$ sufficiently large ensures that $W^{(3)} y + b^{(3)}$ is strictly positive on $(f_2 \circ f_1)(\mathbb{R}^d)$, so the ReLU in the third layer is inactive. Hence, it acts like a linear map. Thus, for $X \in \mathrm{GL}^+_{n_3}$ with small enough norm, we have that replacing the weights with $XW^{(3)}$, $Xb^{(3)}$ and $W^{(4)}X^{-1}$ does not change the function. Hence f_θ is not identifiable. □

F.2 Conserved Quantities from GL-Action on Linear Neurons

We now show that the symmetries presented in Section 4 which are induced by the monoid action of GL^+, give rise to locally conserved quantities. We first make a statement that holds for linear neural networks, and then show that this can be applied to the setting of Proposition 39.

Proposition 50. *Let $\theta = \{(U, V, u, v) \in \mathbb{R}^{m \times J} \times \mathbb{R}^{J \times d} \times \mathbb{R}^m \times \mathbb{R}^J\} \in \Theta$ denote the parameters of a linear network $f : \mathbb{R}^d \to \mathbb{R}^m$ given by $f(x) = U(Vx + v) + u$. For fixed $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$, let $D_{A,b} \subseteq \Theta$ be the subspace given by*

$$D_{A,b} = \left\{(U, V, u, v) \in \mathbb{R}^{m \times J} \times \mathbb{R}^{J \times d} \times \mathbb{R}^m \times \mathbb{R}^J \mid UV = A, \quad Uv + u = b\right\}. \tag{12}$$

Then

$$U^\top U - V V^\top - v v^\top \tag{13}$$

is (locally) preserved on $D_{A,b}$ through gradient flow $\dot{\theta}$.

Proof. Let $X : \mathbb{R} \to \mathrm{GL}_J$ be a curve into the space of invertible matrices. Recall that GL_J acts on $D_{A,b}$ via the group action $X \cdot (U, V, u, v) = (UX^{-1}, XV, u, Xv)$. Then $\gamma(s) = X(s) \cdot \theta$ is a curve into the fiber of a parameter θ. Since $\langle \dot{\gamma}(s), \dot{\theta} \rangle = 0$, a function $C : \Theta \to \mathbb{R}$ with $\nabla_\theta C = \dot{\gamma}(s)$ must be conserved. For simplicity, we will write $Y(s) = X(s)^{-1}$. From

$$\gamma(s) = (UY(s), X(s)V, u, X(s)v) \tag{14}$$

it follows that

$$\dot{\gamma}(s) = (U\dot{Y}(s), \dot{X}(s)V, 0, \dot{X}(s)v). \tag{15}$$

Equation 15 implies that

$$\begin{aligned} \nabla_{V_{ij}} C &= \sum_t \dot{X}_{it} V_{tj}, \\ \nabla_{U_{ij}} C &= \sum_t U_{it} \dot{Y}_{tj}, \\ \nabla_{v_i} C &= \sum_t \dot{X}_t v_t. \end{aligned}$$

Integrating this, we obtain

$$C(\theta) = \frac{1}{2} \cdot \sum_{i,j,t} \left(V_{ji} V_{ti} \dot{X}_{tj} + U_{it} U_{ij} \dot{Y}_{tj} \right) + \frac{1}{2} \cdot \sum_{i,t} \dot{X}_{it} v_t v_i. \tag{16}$$

Let E_{ij} be the $J \times J$ matrix with a 1 in its (i, j)-th entry and zeros elsewhere. Consider now

$$X(s) = I_J + s(E_{mk} + E_{km}), \;\; s \geq 0.$$

This $X(s)$ is invertible for $s \neq \pm 1$, with $\dot{Y}(0) = -(E_{mk} + E_{km})$ and $\dot{X}(0) = E_{mk} + E_{km}$. Plugging these values into (16), we obtain $C(\theta) = (VV^\top)_{km} - (U^\top U)_{km} + (vv^\top)_{km}$. Running through all indices k, m this implies that $VV^\top - U^\top U + vv^\top$. □

Proposition 51. *Let $f_\theta = f^{(1)}_{\theta_1} \circ (f^{(2)}_{\theta_2} + f^{(3)}_{\theta_3}) \circ f^{(4)}_{\theta_4}$ be a neural network (or even just a parameterized function) and assume that on $S \subseteq \Theta_2$ we have a conserved quantity $C_2(\theta_2)$. Then $C((\theta_i)_{i=1,\dots,4}) = C_2(\theta_2)$ is a conserved quantity on $\Theta_1 \times S \times \Theta_3 \times \Theta_4$.*

Proof. Since $C_2(\theta_2)$ is conserved on $S \subseteq \Theta_2$, we have that whenever $\theta_2 \in S$

$$\langle \nabla_{\theta_2} C_2(\theta_2), \dot{\theta}_2 \rangle = 0,$$

where $\dot{\theta}_2 = \pi_2(\dot{\theta})$ is the projection of $\dot{\theta}$ on its second component. Since $\nabla_\theta C(\theta) = (0, \nabla_{\theta_2} C_2(\theta_2), 0, 0)$, it immediately follows that

$$\langle \nabla_\theta C(\theta), \dot{\theta} \rangle = \langle \nabla_{\theta_2} C_2(\theta_2), \dot{\theta}_2 \rangle = 0,$$

whenever $\theta_2 \in S$ or equivalently $\theta \in \Theta_1 \times S \times \Theta_3 \times \Theta_4$. This implies that $C(\theta)$ is conserved under $\dot{\theta}$, as long as $\theta \in \Theta_1 \times S \times \Theta_3 \times \Theta_4$. □

Noting that the curve used in the proof of Proposition 50 is contained in GL^+, we obtain the locally conserved quantity given in Theorem 17.

F.3 Conserved Quantities from Layer Concatenation

Proposition 52. *The continuous symmetry described in Section 5 does not induce a conserved quantity of gradient flow.*

Proof. Consider $T_j(\theta)$ as in Lemma 8. This immediately yields the curve

$$\gamma(t) : (\max_{i \in [n_2]} \frac{b^{(2)}_i}{W^{(2)}_{ij}}, \infty) \to T_j(\theta)$$

given by

$$\gamma(t) = (W^{(1)}, b^{(1)} + te_j, W^{(2)}, b^{(2)} - t(W^{(2)}_{1j}, \dots, W^{(2)}_{n_2 j})^\top, W^{(3)}, b^{(3)}).$$

Since the gradient flow $\dot{\theta} = -\nabla_\theta \mathcal{L}$ is orthogonal to the fiber $\tilde{\mathcal{S}}(\theta)$, we have $\langle \dot{\gamma}(t), \dot{\theta} \rangle = 0$. We want to use this to find a conserved quantity $C : \Theta \to \mathbb{R}$ with

$$\nabla_\theta C = \dot{\gamma}(0). \tag{17}$$

Evaluating $\dot{\gamma}(s)$, we require $\nabla_{W^{(2)}_{i,j}} C = 0$ for all i. However, we have that $\frac{\partial C}{\partial b^{(2)}_i} = -W^{(2)}_{ij}$. Observing that C must have continuous partial derivatives, implying that the mixed partial derivatives must agree, we conclude that there exists no C satisfying (17). Due to the condition $W^{(2)}_{ij} > 0, W^{(2)}_{ik} \leq 0$ for all $k \neq j$ per layer there can exist at most one hidden hyperplane and thus at most one such one-dimensional fiber $T_j(\theta)$. Hence, this symmetry induces no conserved quantities. □

G Discussion: Beyond the Generic Bottleneck Regime

Our main results provide a complete description of generic fibers for three-layer bottleneck architectures. In this appendix, we discuss how the underlying ideas may extend beyond this setting and why explicit descriptions can be expected to become substantially more complicated.

A useful guiding principle is that the symmetry mechanisms identified in this work persist more generally. In particular, both layerwise symmetries and hiding symmetries remain meaningful in wider and deeper architectures. The challenge lies not in their existence, but describing their interactions with other symmetries and in assessing whether they form a complete description.

G.1 Wider Architectures

The bottleneck assumption plays a special geometric role, as it implies that for generic parameters the image of the first hidden layer is the full nonnegative orthant. This reduces the analysis of the second layer to that of a single weighted hyperplane arrangement on a fixed domain.

Once the bottleneck assumption is dropped, this picture changes fundamentally. The image of the first hidden layer is no longer the full orthant, but a union of polyhedra in activation space, with up to $O(n^d)$ such cells in general. To determine which geometric features of the first layer are preserved by the next layer and which are hidden, one must analyze how the next-layer hyperplanes intersect this polyhedral image.

In principle, the same general philosophy should still apply locally on each image cell: one can ask whether a hyperplane of the next layer intersects the relevant part of the image, whether it is visible in the realized function, and whether it constrains the geometry of the previous layers. However, the compact hyperplane representation available in the bottleneck case no longer applies. Instead, one must track many activation patterns separately, and any resulting semi-algebraic description of the fiber is therefore expected to grow combinatorially.

Moreover, wider architectures may introduce additional continuous degrees of freedom that do not arise in the bottleneck case. If the intersection of the image of one layer with a hyperplane of the next fails to span the entire image cell, then the hyperplane may admit deformations that do not change its effect on the image of the previous layers. For instance, one can consider rotations that leave its intersection with the image invariant. In such a regime, the relevant symmetries are no longer directly determined by the weight sign pattern alone.

G.2 Deeper Architectures

For deeper networks, the main difficulty is not only combinatorial. Our non-localizability result (Proposition 16) shows that parameter equivalence cannot, in general, be reduced to pairwise layer compositions. This indicates that, in deeper architectures, one must track the image of an entire network prefix and study how its image is filtered through all subsequent layers.

In other words, explicit fiber descriptions are not expected to be layerwise decomposable. The admissible transformations at a given layer depend on the geometry induced by all preceding layers and on how this geometry is preserved, hidden, or altered by all subsequent layers.

A further difficulty is that breakpoint features may no longer be uniquely attributable to a single layer. In the three-layer bottleneck case, visibility arguments allow us to identify which breakpoint facets come from the first layer and which from the second. In deeper architectures, this full separation may fail. As a result, new discrete symmetries may emerge from different ways of assigning the same realized breakpoint structure to different layers. Such discrete symmetries may also interact nontrivially with continuous symmetries, for example those arising from linear neurons.

G.3 Nongeneric Parameters

The nongeneric case is already substantially more intricate, even for three-layer architectures. The rigidity results used throughout our analysis rely on genericity in two ways: they ensure that hyperplanes intersect in the expected codimensions and prevent algebraic degeneracies in the masked weight products.

Once genericity is dropped, both mechanisms can fail. Hyperplanes may coincide or meet in unexpected dimensions, so that different neurons may contribute indistinguishably to the same geometric feature. In particular, neurons can no longer be assigned unambiguously to visible breakpoints. This opens the door to many further degeneracies, including additional discrete symmetries arising from coincident, canceling, or partially hidden bent hyperplanes.

Thus, while the generic bottleneck case admits an explicit description in terms of visible hyperplanes, orientations, and affine compensation, the nongeneric case is expected to require a substantially richer case distinction.

G.4 Complexity Perspective

The geometric complications described above are supported by complexity-theoretic considerations. In particular, our polynomial-time equivalence test appears to rely essentially on both the bottleneck and genericity assumptions.

First, if the bottleneck assumption is dropped, then deciding functional equivalence of three-layer ReLU networks is already coNP-hard. This follows from hardness results for positivity of one-hidden-layer ReLU networks. For example, Froese et al. (2026) show that, given an instance of MULTICOLORED CLIQUE, one can construct in polynomial time a two-layer ReLU network computing a function $f\colon \mathbb{R}^k \to \mathbb{R}$ such that $\max_{x\in\mathbb{R}^k} f(x) = k + \binom{k}{2}$ in the yes-case, whereas $\max_{x\in\mathbb{R}^k} f(x) \leq k + \binom{k}{2} - 1$ in the no-case. By adding one further ReLU neuron that thresholds the output at the midpoint of this gap, one obtains a three-layer ReLU network whose realized function is identically zero in the no-case and nonzero in the yes-case. Deciding whether this network computes the zero function is therefore coNP-hard, and equivalently so is deciding functional equivalence with the zero network. Since the first hidden layer in this construction can be substantially wider than the input dimension, the resulting hard instances lie outside the bottleneck regime.

Second, hardness persists even within bottleneck architectures if genericity is dropped. The reduction of Froese et al. (2025) for positivity of one-hidden-layer ReLU networks constructs instances in input dimension d with only $O(d^2)$ hidden neurons. Such an instance can be embedded into a larger ambient space of dimension $O(d^2)$ by composing with a projection onto the first d coordinates. This preserves positivity, while the resulting network has first hidden layer width at most the ambient dimension, and is therefore a bottleneck architecture and necessarily nongeneric. In this way, one obtains coNP-hardness of functional equivalence even within the class of bottleneck architectures once genericity is dropped.

Finally, it seems unlikely that restricting to generic parameters alone restores tractability in the non-bottleneck regime. The hardness construction of Froese et al. (2026) is separated by a constant gap, and all potential maximizers lie in a bounded region. This suggests that, after suitably stabilizing the construction outside that region, for example by adding neurons that force the network to attain negative values there, sufficiently small perturbations of the realized function should preserve the distinction between yes- and no-instances. Since generic parameters are dense, this provides strong evidence that coNP-hardness should persist even under a genericity restriction. However, turning this intuition into a formal reduction would require an efficient deterministic procedure for perturbing arbitrary instances into generic ones. At present, we do not know how to construct such perturbations

effectively, since genericity requires avoiding finitely many, but exponentially many, algebraic varieties.

From this perspective, the loss of compact geometric descriptions outside the generic bottleneck setting is not an artifact of our proof technique. Rather, it is consistent with the fact that deciding functional equivalence itself becomes hard in these broader regimes. In particular, one should not expect similarly explicit and efficiently checkable semi-algebraic descriptions of fibers once these assumptions are removed. The combinatorial and geometric blow-up described above may therefore reflect a genuine increase in problem complexity. In this sense, the generic bottleneck setting isolated in this paper appears to lie near a tractability boundary for explicit and efficient descriptions of ReLU network fibers.