Parameter Symmetries Determine Representational Geometry in Overparameterized Nonlinear Networks

Marvin Theiss[†]
University of Tübingen & IMPRS-IS
marvin.theiss@uni-tuebingen.de

Lukas Braun
Allen Institute for Neural Dynamics
lukas.braun@alleninstitute.org

Andrew M. Saxe
Gatsby Unit & SWC, UCL
a.saxe@ucl.ac.uk

Erin Grant
University of Alberta & Amii
eringrant@ualberta.ca

Abstract

Representations are routinely used across machine learning, psychology, and neuroscience to draw inferences about the computations of biological and artificial systems. Such inferences presume that the geometry of these representations reflects the computation being performed. For artificial neural networks, however, the extent to which the function a network computes constrains this geometry remains unclear. One key obstacle is that these networks admit *parameter symmetries*: changes in parameterization that preserve function exactly while reshaping representational geometry. Here, we show that a broad class of parameter symmetries acts on representations through just three primitive feature transformations: addition, duplication, and scaling. This feature-level characterization yields a closed-form decomposition of representational geometry into *essential* and *auxiliary* components, revealing how representational degeneracy can exist without and grows with overparameterization even when function is held fixed. Finally, we establish conditions under which implementation-level selection rules resolve this degeneracy, yielding identifiable geometries in which features are weighted according to their contributions to the network's function. Together, our results delineate when representational geometry can support inferences about computation, and when it cannot.

1 Introduction

What does the geometry of the internal representations of neural networks reveal about the computations these networks implement? Techniques for interpreting and intervening on the internal activities of artificial neural networks assume that local structure, such as weights or activities within a layer, reveals information about the end-to-end computation these networks perform [1, 2, 3, 4]. In neuroscience and cognitive science, representational similarity analysis and related techniques for comparing the internal representations of biological and artificial neural networks share this assumption that the local geometry of neural activity reveals the computations it instantiates, and thus serve as powerful tools for interpreting neural function in terms of neural activity [5, 6, 7].

However, artificial neural networks, like biological neural networks, are degenerate [8, 9], as many distinct parameter configurations [10, 11, 12] and therefore distinct patterns of neural activity [13, 14, 15, 16, 17] can support the same function. Further, Braun et al. [18] analytically established a double dissociation between function and representation in two-layer linear networks: networks can compute identical functions while exhibiting distinct representational geometries, and conversely can

[†]Work partially done while a project research intern at the Gatsby Computational Neuroscience Unit, UCL.

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

exhibit identical representational geometries while computing distinct functions. If such a double dissociation were general, it would call into question whether representational geometry can serve as a suitable basis for studying computation at all. Determining whether this double dissociation persists in *nonlinear* neural networks is therefore consequential both practically, because modern applications rely on nonlinear networks, and theoretically, because a given function admits a wider range of representations while the space of realizable functions is itself more complex.

Here, we study how strongly *parameter symmetries*, transformations of a network's parameters that leave the realized function unchanged [19, 20, 21, 22, 23], can dissociate representation from function in nonlinear networks. By formalizing and refining parameter symmetries described in prior work [12, 24] into a precise taxonomy, and deriving their exact effects on representations, we prove that these symmetries act through compositions of just three primitive feature transformations: *addition*, *duplication*, and *scaling*. This characterization exposes substantial variability in representational geometry, demonstrating that functional equivalence need not imply representational alignment.

Yet this feature-transformation view also reveals a path to identifiability: suitable implementation-level constraints remove symmetry-induced degrees of freedom and restore unique representational geometries, yielding a nonlinear analog of the linear result [18]. These constraints select parameterizations that calibrate features by computational importance, thereby identifying privileged representational geometries even within the degenerate class of nonlinear neural networks. Together, our results delineate when representational geometry can support inferences about computation and when function-preserving variation in implementation renders such inferences fundamentally ambiguous.

Concretely, our main contributions are as follows:

- **Symmetry-equivalence**. We formalize and refine parameter symmetries in nonlinear one-hidden-layer networks described in prior work into a precise taxonomy, and derive a necessary and sufficient criterion for when two parameterizations are symmetry-equivalent (Section 3).
- **Three feature primitives**. We show that parameter symmetries act on hidden representations through feature addition, duplication, and scaling, and characterize hidden-activation matrices within a symmetry orbit (Section 4).
- **Representational degeneracy**. We decompose representational geometry into essential and auxiliary components, explaining why and characterizing how degeneracy grows with overparameterization as function stays fixed (Section 5).
- **Representational identifiability**. We establish conditions under which selection by minimum norm restores representational identifiability and show that the selected parameterizations weight features according to their contributions to the realized function (Section 6).

The code used to generate all figures and experimental results reported in this paper is publicly available at github.com/mrvnthss/symmetries-representational-geometry.

2 Preliminaries and setting

Nonlinear networks. We consider nonlinear, fully connected one-hidden-layer networks with incoming weights $\mathbf{W} \in \mathbb{R}^{N_h \times N_i}$, hidden biases $\mathbf{b} \in \mathbb{R}^{N_h}$, and readout weights $\mathbf{A} \in \mathbb{R}^{N_o \times N_h}$. Let $\mathbf{w}_j^\top$, b_j, and $\mathbf{a}_j$ denote the jth row, entry, and column of $\mathbf{W}$, $\mathbf{b}$, and $\mathbf{A}$, respectively, and define the augmented vectors $\overline{\mathbf{w}}_j^\top := (\mathbf{w}_j^\top, b_j)$ and $\overline{\mathbf{x}}^\top := (\mathbf{x}^\top, 1)$. The network then computes

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^{N_h} \mathbf{a}_j \, \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) = \sum_{j=1}^{N_h} \mathbf{a}_j \, \sigma(\overline{\mathbf{w}}_j^\top \overline{\mathbf{x}}), \qquad \mathbf{x} \in \mathbb{R}^{N_i}, \tag{1}$$

where $\sigma \colon \mathbb{R} \to \mathbb{R}$ is the nonlinearity and $\boldsymbol{\theta} := (\sigma; \mathbf{W}, \mathbf{b}, \mathbf{A})$ denotes the parameterization. We assume that σ is nonaffine, i.e., there exist no $m, c \in \mathbb{R}$ such that $\sigma(z) = mz + c$ for every $z \in \mathbb{R}$.

Hidden activations. Given inputs $\mathbf{x}^\mu \in \mathbb{R}^{N_i}$, $\mu = 1, \dots, P$, let $\mathbf{h}^\mu \in \mathbb{R}^{N_h}$ denote the corresponding hidden activation. We collect inputs and hidden activations as

$$\mathbf{X} := [\mathbf{x}^1, \dots, \mathbf{x}^P] \in \mathbb{R}^{N_i \times P}, \qquad \mathbf{H} := [\mathbf{h}^1, \dots, \mathbf{h}^P] \in \mathbb{R}^{N_h \times P}. \tag{2}$$

For later use, define the augmented matrix $\overline{\mathbf{X}} := [\overline{\mathbf{x}}^1, \dots, \overline{\mathbf{x}}^P]$. The hidden-activation matrix $\mathbf{H}$ admits two complementary interpretations: its μth column, $\mathbf{h}^\mu$, is the population response to input $\mathbf{x}^\mu$, while its jth row, $\mathbf{v}_j^\top \in \mathbb{R}^{1 \times P}$, records the activity of neuron j across inputs.

Representational geometry. The hidden-activation matrix $\mathbf{H}$ gives rise to the uncentered Gram matrix $\mathbf{H}^\top\mathbf{H} \in \mathbb{R}^{P\times P}$, whose (μ,ν)th entry is the inner product between the population responses elicited by inputs $\mathbf{x}^\mu$ and $\mathbf{x}^\nu$ [25, 26]. This matrix, known as the *representational similarity matrix (RSM)*, is a widely used descriptor of representational geometry and a natural object for studying variation therein, since many measures of representational similarity depend on $\mathbf{H}$ only through $\mathbf{H}^\top\mathbf{H}$ [6, 27, 28]. The two interpretations of $\mathbf{H}$ in the preceding paragraph yield complementary views of the RSM: at the population level, it encodes pairwise similarities between input-evoked responses, which is the usual interpretation; at the neuron level, the RSM decomposes into rank-one contributions from individual neurons:[2]

$$\mathbf{H}^\top\mathbf{H} = \sum_{j=1}^{N_h} \mathbf{v}_j\mathbf{v}_j^\top, \qquad \mathbf{v}_j^\top = \sigma(\overline{\mathbf{w}}_j^\top\overline{\mathbf{X}}) \in \mathbb{R}^{1\times P}. \tag{3}$$

This neuron-wise decomposition is particularly suited to our analysis: by isolating each hidden neuron's contribution, it allows us to track how these contributions change under function-preserving transformations of the network's parameterization $\boldsymbol{\theta}$. These transformations are precisely the parameter symmetries considered in the next section.

3 Functional parameter symmetries in nonlinear networks

A given function f may be realized by many distinct parameterizations, each inducing its own hidden-activation matrix and corresponding RSM. Here, we study the resulting variation in representational geometry by considering parameterizations related by *parameter symmetries*, transformations of a network's parameters that leave a specified quantity unchanged. Such symmetries admit several definitions, differing in both the quantity required to remain invariant and the set of inputs over which invariance is imposed [23]. We focus throughout on the strictest notion, *functional parameter symmetries*, which leave the network function in Equation (1) unchanged over the entire input domain. In this section, we introduce the symmetries we consider, define symmetry orbits and the related notions of irreducibility and overparameterization, and derive orbit invariants that provide a necessary and sufficient criterion for determining whether two parameterizations are symmetry-equivalent.

3.1 A catalog of parameter symmetries

The parameter symmetries we consider take qualitatively distinct forms. Some act on *individual* neurons, including the positive-scaling symmetry of positively 1-homogeneous activations such as ReLU [22], which rescales a neuron's incoming parameters and readout weights by reciprocal factors, and the sign-flip symmetry of odd activations such as tanh [19]. Others exploit redundancy among *groups* of neurons, including *zero groups*, in which neurons cancel one another through their readouts, and *duplicate groups*, whose copies collectively reproduce the computation of a single neuron [12]. A third class of symmetries couples groups of neurons through algebraic properties of the activation σ, such as *linear groups* that can arise when σ decomposes into the sum of an even function and a linear function, as do ReLU and several other commonly used activations [24]. Panel A of Figure 1 illustrates all parameter symmetries we consider, building on and refining the taxonomies of Şimşek et al. [12] and Martinelli et al. [24], with detailed definitions deferred to Appendix B.

The algebraic conditions underlying this third class of parameter symmetries can be characterized by whether σ is *even-linear*, $\sigma(z) = e(z) + mz$ with e even; *constant-odd*, $\sigma(z) = c + o(z)$ with o odd; or neither [12, 24]. This classification includes positively 1-homogeneous activations, such as ReLU, which form a subclass of even-linear activations (Corollary C.2). Table C.1 classifies other commonly used activations by the symmetries they admit, with detailed derivations in Appendix C.

3.2 Symmetry orbits, irreducibility, and overparameterization

The functional parameter symmetries described in Section 3.1 naturally generate an equivalence relation on network parameterizations, including those of different widths. We call two parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ *symmetry-equivalent*, and write $\boldsymbol{\theta} \sim \boldsymbol{\xi}$, if they are connected by a finite, function-preserving composition of the parameter symmetries enumerated in Appendices B.1 to B.3.

[2] Since neuron permutations leave $\mathbf{H}^\top\mathbf{H}$ unchanged, we understand all identities involving hidden-activation matrices $\mathbf{H}$, like that in Equation (3), up to a permutation of the rows of $\mathbf{H}$, with all neuron-indexed quantities relabeled consistently.

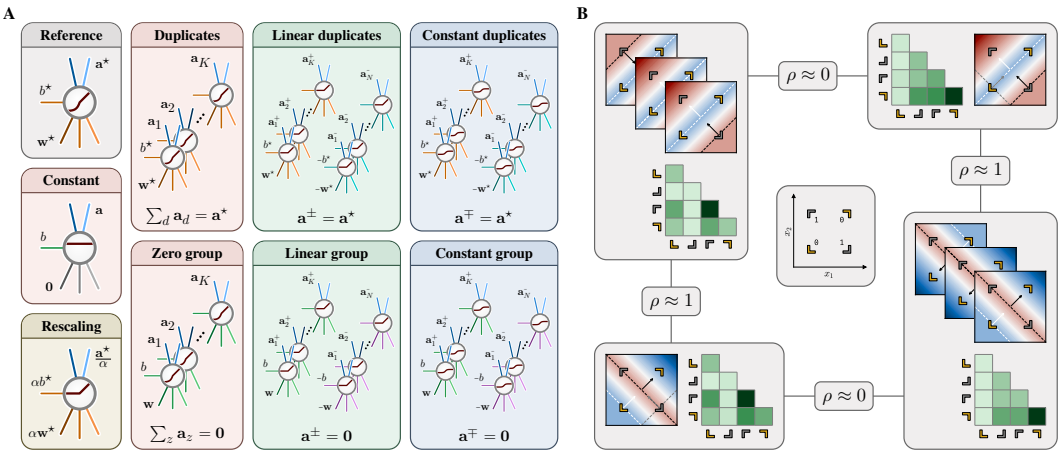

Figure 1: **Parameter symmetries can doubly dissociate function and representation**. **(A)** Parameter symmetries let networks implement the same function via different parameterizations. They act on *individual* neurons (e.g., positive scaling for ReLU networks), redistribute or cancel computation across *groups* (duplicates, zero groups), or couple groups of neurons through algebraic structure in the activation σ (linear or constant duplicates, linear or constant groups). **(B)** Gradient descent from random initializations yields six distinct two-neuron ReLU networks solving XOR (center), shown as input-space heatmaps with neuron decision boundaries overlaid. They form two families (top left and bottom right) that solve the same task using different functions, with RSMs of solutions from opposite families being nearly uncorrelated. Crucially, such representational variability persists even when the function is held fixed: overparameterized networks implementing *exactly the same* function can have nearly uncorrelated RSMs (top and bottom). Conversely, networks implementing *different* functions can have highly correlated RSMs (left and right), yielding a double dissociation.

Definition 3.1 (Symmetry orbit)**.** The *symmetry orbit* $\mathcal{O}(\boldsymbol{\theta})$ of a parameterization $\boldsymbol{\theta}$ is the set $\mathcal{O}(\boldsymbol{\theta}) \coloneqq \{\boldsymbol{\xi} \mid \boldsymbol{\xi} \sim \boldsymbol{\theta}\}$ consisting of all parameterizations $\boldsymbol{\xi}$ that are symmetry-equivalent to $\boldsymbol{\theta}$. The *orbit at width* N_h is the subset $\mathcal{O}_{N_h}(\boldsymbol{\theta}) \coloneqq \{\boldsymbol{\xi} \in \mathcal{O}(\boldsymbol{\theta}) \mid \boldsymbol{\xi} \text{ has width } N_h\} \subseteq \mathcal{O}(\boldsymbol{\theta})$.

Minimum-width parameterizations within a symmetry orbit play a central role in our analysis.

Definition 3.2 (Irreducibility and overparameterization)**.** A parameterization $\boldsymbol{\theta}$ is called *irreducible* if no other element of its symmetry orbit $\mathcal{O}(\boldsymbol{\theta})$ has smaller width, and *overparameterized* otherwise.

Irreducible parameterizations within an orbit need *not* be unique. Neuron permutations generate finite families of irreducible parameterizations. When admitted by the activation σ, sign flips and positive scaling generate additional irreducible parameterizations, with the latter producing continuous families thereof. Perhaps counterintuitively, an orbit can contain distinct irreducible parameterizations that cannot be related by any combination of these symmetries alone (Example B.13).

3.3 Orbit invariants and essential parameter classes

The definition of symmetry-equivalence just introduced is not directly operational. Determining whether two parameterizations are equivalent requires either explicitly constructing a sequence of symmetries that transforms one into the other or proving that no such sequence exists. Here, we establish a necessary and sufficient criterion characterizing symmetry-equivalence for positively homogeneous activations of degree 1, with all remaining activation classes treated in Appendix B.4.

By Corollary C.2, every nonaffine positively 1-homogeneous activation σ can be expressed as $\sigma(z) = \delta|z| + mz$ with $\delta \neq 0$ and $m \in \mathbb{R}$. Hence, the contribution of a nonconstant neuron with parameters $(\overline{\mathbf{w}}, \mathbf{a})$ can be decomposed into nonlinear and linear components

$$\|\overline{\mathbf{w}}\|\mathbf{a}\left(\delta \frac{|z|}{\|\overline{\mathbf{w}}\|}\right) \qquad \text{and} \qquad \mathbf{a}mz, \tag{4}$$

respectively, where $z \coloneqq \overline{\mathbf{w}}^\top \overline{\mathbf{x}}$. The nonlinear part depends on the neuron's parameterization only through the activation hyperplane defined by $\overline{\mathbf{w}}^\top \overline{\mathbf{x}} = 0$ and the rescaled readout $\|\overline{\mathbf{w}}\|\mathbf{a}$. We therefore

identify incoming-parameter vectors up to nonzero scaling, defining the *parameter class* of $\overline{\mathbf{w}}$ as $[\overline{\mathbf{w}}] \coloneqq \{\alpha\overline{\mathbf{w}} \mid \alpha \in \mathbb{R}_{\neq 0}\}$, and we collect the classes represented by $\boldsymbol{\theta}$ in $\mathcal{Q}(\boldsymbol{\theta}) \coloneqq \{[\overline{\mathbf{w}}_j] \mid \mathbf{w}_j \neq \mathbf{0}\}$. Let $\mathcal{J}_q \coloneqq \{j \mid \overline{\mathbf{w}}_j \in q\}$ index the neurons of class $q \in \mathcal{Q}(\boldsymbol{\theta})$, and let $\mathcal{J}_0 \coloneqq \{j \mid \mathbf{w}_j = \mathbf{0}\}$ index all constant neurons. For each class $q = [\overline{\mathbf{w}}] \in \mathcal{Q}(\boldsymbol{\theta})$, we define the *normalized nonlinear feature* ϕ_q shared by all neurons in q and the corresponding *aggregate rescaled readout* $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ as

$$\phi_q(\mathbf{x}) \coloneqq \delta \frac{|z|}{\|\overline{\mathbf{w}}\|}, \qquad \boldsymbol{\beta}_q(\boldsymbol{\theta}) \coloneqq \sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\| \, \mathbf{a}_j, \tag{5}$$

respectively. Finally, we collect the linear residual $\mathbf{a}_j m z_j$ of nonconstant neurons and the total contribution $\mathbf{a}_j \sigma(b_j)$ of constant neurons into the map $\mathbf{r}_{\boldsymbol{\theta}} \colon \mathbb{R}^{N_i} \to \mathbb{R}^{N_o}$ given by

$$\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq m \sum_{j \notin \mathcal{J}_0} \mathbf{a}_j z_j + \sum_{j \in \mathcal{J}_0} \mathbf{a}_j \sigma(b_j), \qquad z_j \coloneqq \overline{\mathbf{w}}_j^\top \overline{\mathbf{x}}. \tag{6}$$

With this notation in place, the function $f_{\boldsymbol{\theta}} \colon \mathbb{R}^{N_i} \to \mathbb{R}^{N_o}$ realized by $\boldsymbol{\theta}$ can be expressed as

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{q \in \mathcal{Q}(\boldsymbol{\theta})} \boldsymbol{\beta}_q(\boldsymbol{\theta}) \phi_q(\mathbf{x}) + \mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}). \tag{7}$$

Proposition 3.3 (Characterization of symmetry-equivalence)**.** *Two parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, possibly of different widths, with the same nonaffine positively homogeneous activation of degree* 1*, are symmetry-equivalent if and only if*

$$\boldsymbol{\beta}_q(\boldsymbol{\theta}) = \boldsymbol{\beta}_q(\boldsymbol{\xi}) \quad \text{for every } q \in \mathcal{Q}(\boldsymbol{\theta}) \cup \mathcal{Q}(\boldsymbol{\xi}), \qquad \mathbf{r}_{\boldsymbol{\theta}} = \mathbf{r}_{\boldsymbol{\xi}}, \tag{8}$$

with $\boldsymbol{\beta}_q(\cdot)$ understood to be zero for an absent class. (Proof in Appendix B.4.2.)

Within a fixed orbit $\mathcal{O}$, we suppress the parameterization argument $\boldsymbol{\theta}$ and call a class q *essential* if $\boldsymbol{\beta}_q \neq \mathbf{0}$ and *nonessential* otherwise. We write $\mathcal{E} \coloneqq \{q \mid \boldsymbol{\beta}_q \neq \mathbf{0}\}$ for the set of essential classes.

4 Parameter symmetries act through three feature primitives

While two symmetry-equivalent parameterizations realize the same function, they can induce drastically different hidden activations. Overparameterization amplifies this variability, as additional hidden neurons create increasingly many ways to compose parameter symmetries and distribute computation across neurons without ever changing the realized function. This raises the question of whether the corresponding hidden-activation matrices admit a tractable description. The key observation of this section is that they do: at the level of hidden activations, arbitrary sequences of parameter symmetries reduce to compositions of just three primitive feature transformations. Together with the orbit invariants of Section 3.3, this yields an explicit characterization of hidden-activation matrices within a symmetry orbit. Up to scaling and duplication, each essential parameter class remains represented by at least one feature throughout the orbit, while nonessential classes may contribute additional features. For several important activation classes, persistence throughout the orbit extends from essential parameter classes to the individual features shared by its irreducible parameterizations.

4.1 Feature-level action of parameter symmetries

The parameter symmetries introduced in Section 3 take several distinct forms in parameter space: they may introduce canceling groups of neurons, duplicate existing neurons, or exploit algebraic structure of the activation through sign-flipped incoming parameters. At the level of hidden activations, however, these distinct parameter-space mechanisms reduce to compositions of just three primitive feature transformations, which we introduce next.

Feature *addition* appends features computed by additional neurons. Collecting these features in $\mathbf{U} \in \mathbb{R}^{K \times P}$, we define $\mathcal{A}_{\mathbf{U}}(\mathbf{H}) \coloneqq [\mathbf{H}^\top, \mathbf{U}^\top]^\top$. Feature *duplication* repeats existing features. Given a duplication pattern $\boldsymbol{\nu} \in \mathbb{N}_{>0}^{N_h}$, let $\mathbf{D}_{\boldsymbol{\nu}}$ denote the duplication matrix that repeats row i exactly ν_i times (see Definition D.2), and define $\mathcal{D}_{\boldsymbol{\nu}}(\mathbf{H}) \coloneqq \mathbf{D}_{\boldsymbol{\nu}}\mathbf{H}$. Feature *scaling* multiplies each feature by a nonzero scalar. For $\boldsymbol{\alpha} \in \mathbb{R}_{\neq 0}^{N_h}$, we define $\mathcal{S}_{\boldsymbol{\alpha}}(\mathbf{H}) \coloneqq \operatorname{diag}(\boldsymbol{\alpha})\mathbf{H}$.

The corresponding inverse operations remove appended features, discard all but one copy of duplicated feature rows, and apply reciprocal scaling factors, respectively. Together, these operations provide an exhaustive characterization of how parameter symmetries act on hidden activations.

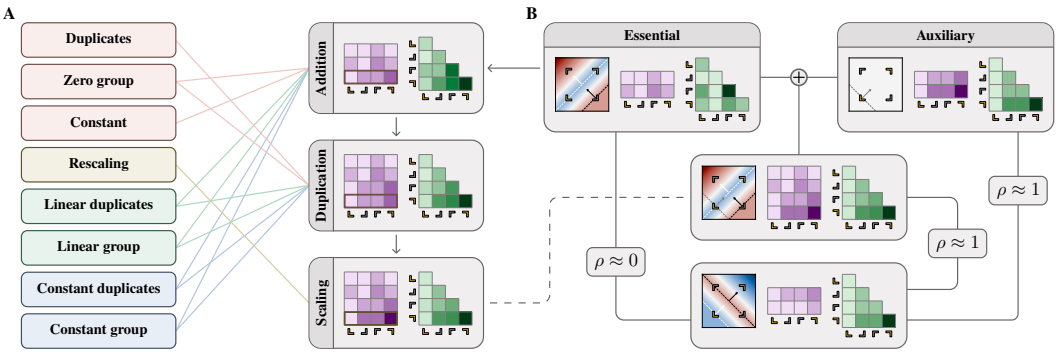

Figure 2: **Primitive feature transformations unify parameter symmetries and separate essential from auxiliary geometry**. **(A)** Parameter symmetries decompose into three primitive feature transformations and their inverses: *addition*, *duplication*, and *scaling*. Shown are their effects on hidden activations and RSMs when adding a zero-readout neuron to a solution in Panel B of Figure 1, duplicating it, and rescaling both copies. **(B)** Decomposing the full RSM of the resulting network (center) into essential and auxiliary contributions reveals the source of the dissociation in Figure 1. Whereas the essential component is nearly uncorrelated with the RSM of an opposite-family solution, the auxiliary component is nearly perfectly correlated with it, driving the full RSM toward the reference geometry without changing the function being computed.

Proposition 4.1 (Feature-level characterization of parameter symmetries)**.** *Up to row permutations, positive-scaling and sign-flip symmetries act through feature scaling, while duplicate-neuron groups act through feature duplication. Zero-neuron groups and constant neurons act through feature addition followed, where necessary, by duplication. Linear-neuron and constant-neuron groups add two oppositely oriented features and duplicate them according to the sizes of their aligned and opposite subgroups. Linear-duplicate and constant-duplicate groups act analogously, except that one orientation is already supplied by the reference neuron. Reversing any of these parameter symmetries acts through the corresponding inverse primitives.* (Proof in Appendix D.2.)

4.2 Structure of hidden-activation matrices within a symmetry orbit

Having characterized individual parameter symmetries at the feature level, we now derive a structural form satisfied by hidden-activation matrices across an entire symmetry orbit. As in Section 3.3, we focus on positively 1-homogeneous activations in the main text and defer the remaining activation classes to Appendix D.3. Fix a symmetry orbit, and let $\mathcal{E}$ denote its essential parameter classes. For each essential class $q \in \mathcal{E}$, choose a unit-norm representative $\overline{\mathbf{w}}_q \in q$, and let $\mathbf{F}_{\mathcal{E}} \in \mathbb{R}^{2|\mathcal{E}| \times P}$ collect the two features

$$(\mathbf{f}_q^+)^\top := \sigma(\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}}), \qquad (\mathbf{f}_q^-)^\top := \sigma(-\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}}). \tag{9}$$

Proposition 4.2 (Hidden-activation matrices within a symmetry orbit)**.** *Fix a symmetry orbit $\mathcal{O}$ with a positively 1-homogeneous activation and essential parameter classes $\mathcal{E}$, and let $\mathbf{F}_{\mathcal{E}}$ be defined as above. Every parameterization $\boldsymbol{\theta}$ of width N_h in $\mathcal{O}$ has a hidden-activation matrix of the form*

$$\mathbf{H} = \operatorname{diag}(\boldsymbol{\alpha}) \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{F}_{\mathcal{E},\mathcal{I}} \\ \mathbf{U} \end{bmatrix}, \qquad \boldsymbol{\alpha} \in \mathbb{R}_{>0}^{N_h}, \quad \boldsymbol{\nu} \in \mathbb{N}_{>0}^{|\mathcal{I}|+K}, \tag{10}$$

for some $K \geq 0$ and $\mathbf{U} \in \mathbb{R}^{K \times P}$, with $\sum_\ell \nu_\ell = N_h$. Here, $\mathbf{F}_{\mathcal{E},\mathcal{I}}$ is the submatrix of $\mathbf{F}_{\mathcal{E}}$ indexed by $\mathcal{I} \subseteq \{1, \ldots, 2|\mathcal{E}|\}$. The index set $\mathcal{I}$ selects the rows of $\mathbf{F}_{\mathcal{E}}$ corresponding to the features represented in $\boldsymbol{\theta}$, with at least one of $(\mathbf{f}_q^+)^\top$ and $(\mathbf{f}_q^-)^\top$ selected for every $q \in \mathcal{E}$. The matrix $\mathbf{U}$ collects pairwise distinct features from nonessential classes and constant neurons. (Proof in Appendix D.3.)

Every parameterization in the same symmetry orbit must therefore represent each essential parameter class by at least one hidden neuron. An even stronger result holds for the following three activation classes: activations that are neither even-linear nor constant-odd (e.g., those studied in Şimşek et al. [12]), even activations (e.g., Gaussian), and odd activations (e.g., tanh). For these classes, persistence extends from essential parameter classes to the individual features themselves. That is, every feature computed by an irreducible parameterization persists throughout the orbit up to nonzero scaling.

Corollary 4.3. *Suppose that σ is neither even-linear nor constant-odd, or is purely even or purely odd. Let $\boldsymbol{\theta}^\star$ be any irreducible parameterization of width $N_h^\star$ with hidden-activation matrix $\mathbf{H}^\star$. Then, every parameterization $\boldsymbol{\theta} \sim \boldsymbol{\theta}^\star$ of width N_h has a hidden-activation matrix of the form*

$$\mathbf{H} = \operatorname{diag}(\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{U} \end{bmatrix}, \qquad \boldsymbol{\alpha} \in \mathbb{R}^{N_h}_{\neq 0}, \quad \boldsymbol{\nu} \in \mathbb{N}^{N_h^\star + K}_{>0}, \tag{11}$$

for some $K \geq 0$ and $\mathbf{U} \in \mathbb{R}^{K \times P}$, with $\sum_\ell \nu_\ell = N_h$. The matrix $\mathbf{U}$ can be chosen to have pairwise distinct rows. (Proof in Appendix D.4.)

5 Parameter symmetries dissociate function and representation

The feature-level characterization in Section 4 lets us translate variability in individual features across a symmetry orbit into variability of the representational geometry these features collectively induce. While the function realized by a parameterization remains unchanged, feature addition can introduce new features into its hidden representation, while duplication and scaling can alter the relative contributions of existing ones. These degrees of freedom are especially consequential because similarity to external reference geometries is often used to draw inferences about the computations supported by a representation, for example when comparing network representations with neural data. We therefore characterize how the range of similarities to a fixed reference geometry attainable within a symmetry orbit expands with overparameterization, eventually reaching bounds that are independent of the realized function.

5.1 Essential and auxiliary contributions to representational geometry

Substituting the factorization of the hidden-activation matrix $\mathbf{H}$ established in Section 4.2 into $\mathbf{H}^\top\mathbf{H}$ yields a decomposition of representational geometry into essential and auxiliary contributions.

Proposition 5.1 (RSM decomposition)**.** *For a parameterization $\boldsymbol{\theta}$ with positively 1-homogeneous activation, let $\mathbf{f}_\ell^\top$ and $\mathbf{u}_k^\top$ denote the rows of $\mathbf{F}_{\mathcal{E},\mathcal{I}}$ and $\mathbf{U}$, respectively, in the factorization of $\mathbf{H}$ in Proposition 4.2. Then,*

$$\mathbf{H}^\top\mathbf{H} = \underbrace{\sum_{\ell=1}^{|\mathcal{I}|} \gamma_\ell \mathbf{f}_\ell \mathbf{f}_\ell^\top}_{\text{essential}} + \underbrace{\sum_{k=1}^{K} \gamma_{|\mathcal{I}|+k} \mathbf{u}_k \mathbf{u}_k^\top}_{\text{auxiliary}}, \qquad \boldsymbol{\gamma} \coloneqq \mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2 \in \mathbb{R}^{|\mathcal{I}|+K}_{>0}, \tag{12}$$

where $\boldsymbol{\alpha}^2$ is the Hadamard square of $\boldsymbol{\alpha}$. For any such realized factorization, every strictly positive $\boldsymbol{\gamma}$ is realizable by a symmetry-equivalent parameterization of the same width, with the feature matrices $\mathbf{F}_{\mathcal{E},\mathcal{I}}$ and $\mathbf{U}$ and duplication pattern $\boldsymbol{\nu}$ fixed. (Proof in Appendix E.1.)

This decomposition reveals two complementary mechanisms by which parameter symmetries reshape representational geometry. First, feature addition can introduce new auxiliary features and hence new rank-one contributions to the geometry $\mathbf{H}^\top\mathbf{H}$, provided that the class aggregates $\boldsymbol{\beta}_q$ and global residual $\mathbf{r}$ are preserved (Section 3.3). Second, duplication and scaling reweight existing contributions. Duplication does so through integer feature multiplicities and is available even without the positive-scaling symmetry, whereas positive scaling permits *arbitrary* strictly positive weights on the rank-one contributions $\mathbf{f}_\ell\mathbf{f}_\ell^\top$ and $\mathbf{u}_k\mathbf{u}_k^\top$. Additional width can therefore accommodate more auxiliary features and a broader range of multiplicities, potentially expanding the set of compatible representational geometries. We next show how this width-dependent freedom in representational geometry translates into an expanding range of similarity scores to a fixed reference geometry.

5.2 Variability in representational similarity

Fix a symmetry orbit $\mathcal{O}$ and denote the RSM induced by a parameterization $\boldsymbol{\theta} \in \mathcal{O}$ on the inputs $\mathbf{X}$ by $\mathbf{M}_{\boldsymbol{\theta}} \coloneqq \mathbf{H}^\top\mathbf{H}$. Let $\mathbf{N} \in \mathbb{R}^{P \times P}$ denote a reference RSM computed from responses to the same inputs, such as population responses measured by fMRI, EEG, or neural recordings. We assume throughout that $\mathbf{N}$ has nonconstant strict upper-triangular entries. Following common practice [6], we quantify alignment with this reference geometry by the Pearson correlation $\rho(\mathbf{M}_{\boldsymbol{\theta}}, \mathbf{N})$ between the strict upper-triangular entries of $\mathbf{M}_{\boldsymbol{\theta}}$ and $\mathbf{N}$.

Let $\mathcal{R}$ denote the realizable RSMs within $\mathcal{O}$ whose strict upper-triangular entries are nonconstant:

$$\mathcal{R} \coloneqq \{\mathbf{M}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{O},\ \mathbf{M}_{\boldsymbol{\theta}} \text{ has nonconstant strict upper-triangular entries}\}. \tag{13}$$

For each width N_h, let $\mathcal{R}_{N_h} \subseteq \mathcal{R}$ denote the subset of RSMs realized in $\mathcal{O}_{N_h}$. The attainable similarity scores across all widths and at width N_h, respectively, are the sets

$$\mathcal{S}(\mathbf{N}) \coloneqq \{\rho(\mathbf{M}_{\boldsymbol{\theta}}, \mathbf{N}) \mid \mathbf{M}_{\boldsymbol{\theta}} \in \mathcal{R}\}, \qquad \mathcal{S}_{N_h}(\mathbf{N}) \coloneqq \{\rho(\mathbf{M}_{\boldsymbol{\theta}}, \mathbf{N}) \mid \mathbf{M}_{\boldsymbol{\theta}} \in \mathcal{R}_{N_h}\}, \tag{14}$$

which we assume to be nonempty. The spreads

$$\Delta_{N_h}(\mathbf{N}) \coloneqq \sup \mathcal{S}_{N_h}(\mathbf{N}) - \inf \mathcal{S}_{N_h}(\mathbf{N}), \qquad \Delta_{\infty}(\mathbf{N}) \coloneqq \sup \mathcal{S}(\mathbf{N}) - \inf \mathcal{S}(\mathbf{N}) \tag{15}$$

quantify the variability in attainable similarity scores at width N_h and across widths, respectively.

Proposition 5.2 (Strict growth until saturation)**.** *For a symmetry orbit with positively* 1*-homogeneous activation, the realizable RSM sets and attainable similarity sets are nested with increasing width:*

$$\mathcal{R}_{N_h} \subseteq \mathcal{R}_{N_h+1}, \qquad \mathcal{S}_{N_h}(\mathbf{N}) \subseteq \mathcal{S}_{N_h+1}(\mathbf{N}). \tag{16}$$

Moreover, $\Delta_{N_h+1}(\mathbf{N}) > \Delta_{N_h}(\mathbf{N})$ *if and only if* $\Delta_{N_h}(\mathbf{N}) < \Delta_{\infty}(\mathbf{N})$. (Proof in Appendix E.2.)

Hence, as overparameterization increases, the spread of similarity scores compatible with the same realized function increases strictly until its maximum within the symmetry orbit is reached. More strikingly, this limit is *independent* of the function being computed: the limiting similarity bounds depend only on the activation σ, probe inputs $\mathbf{X}$, and reference geometry $\mathbf{N}$.

Proposition 5.3 (Function-independent similarity limits)**.** *For fixed positively* 1*-homogeneous activation* σ*, probe inputs* $\mathbf{X}$*, and reference geometry* $\mathbf{N}$*, the limiting similarity bounds* $\inf \mathcal{S}(\mathbf{N})$ *and* $\sup \mathcal{S}(\mathbf{N})$ *are identical across all symmetry orbits* $\mathcal{O}$*, regardless of the functions realized by their parameterizations. Moreover, within each orbit, these common bounds are recovered at sufficiently large finite width. That is, for all sufficiently large* N_h*,*

$$\inf \mathcal{S}_{N_h}(\mathbf{N}) = \inf \mathcal{S}(\mathbf{N}), \qquad \sup \mathcal{S}_{N_h}(\mathbf{N}) = \sup \mathcal{S}(\mathbf{N}). \tag{17}$$

Consequently, every orbit reaches the same limiting spread $\Delta_{\infty}(\mathbf{N})$ *at finite width.* (Proof in Appendix E.3.)

Networks computing entirely different functions can thus approach the same bounds on similarity to a fixed reference geometry through function-preserving reparameterizations.

When $\operatorname{rank}(\overline{\mathbf{X}}) = P$, the attainable similarity range reaches its maximal possible extent. That is, for every reference geometry $\mathbf{N}$, correlations arbitrarily close to both 1 and -1 are attainable within every symmetry orbit (Proposition E.8). For $P \leq N_i + 1$, this rank condition holds generically, making this case particularly relevant for networks probed with high-dimensional inputs such as images. Moreover, these limiting bounds are reached at every width $N_h \geq N_h^{\star} + \frac{P(P-1)}{2} - 1$, where $N_h^{\star}$ denotes the minimum width within the symmetry orbit.

Another important case, which does not require this rank condition, arises when the reference geometry $\mathbf{N}$ is itself the RSM of a network with the same activation σ evaluated on the same inputs $\mathbf{X}$. In this case, the common upper similarity limit is 1, even if the reference network computes an entirely different function (Proposition E.7). The features of the reference network can be introduced through zero-neuron groups, and their rank-one contributions amplified through duplication and scaling until they dominate the RSM, all while preserving the realized function. The construction in Figure 2 illustrates this mechanism with a single feature from an opposite-family solution in Figure 1, whose rank-one contribution is nearly perfectly correlated with the reference RSM.

6 Identifiability through minimum-norm selection

The preceding results show that function alone leaves substantial freedom in representational geometry. Within a single symmetry orbit, the range of attainable similarity scores grows systematically with width and eventually reaches function-independent limits. When the probe inputs are affinely independent, the same realized function is compatible with representations ranging from near-perfect alignment to near-perfect anti-alignment with any reference geometry. Together, these results pose a fundamental obstacle to representational identifiability. Here, however, we show that this

degeneracy can be resolved by accounting for how a parameterization is selected within a symmetry orbit. Specifically, we derive conditions under which minimum-norm selection yields a unique representational geometry, independently of width.

Following the minimum-norm selection rules studied by Braun et al. [18] in linear networks, we consider two criteria for selecting among symmetry-equivalent parameterizations. Within a fixed-width orbit $\mathcal{O}_{N_h}$, we call a parameterization $\boldsymbol{\theta} = (\sigma; \mathbf{W}, \mathbf{b}, \mathbf{A})$ a *minimum-weight-norm parameterization (MWNP)* or *minimum-representation-norm parameterization (MRNP)* if it minimizes, respectively,

$$\Omega_W(\boldsymbol{\theta}) := \|\mathbf{W}\|_F^2 + \|\mathbf{b}\|^2 + \|\mathbf{A}\|_F^2, \qquad \Omega_H(\boldsymbol{\theta}) := \|\mathbf{H}\|_F^2 + \|\mathbf{A}\|_F^2. \tag{18}$$

Both criteria balance the cost of computing features against the cost of reading them out, albeit in slightly different ways. As in previous sections, we focus on positively 1-homogeneous activations and, here, on MWNPs. The corresponding MRNP analysis is deferred to Appendix F.

Fix a symmetry orbit with positively 1-homogeneous activation $\sigma(z) = \delta|z| + mz$, and let $\mathbf{f}_q^+$ and $\mathbf{f}_q^-$ denote the feature vectors generated by the two orientations of the unit-norm representative $\overline{\mathbf{w}}_q$ of an essential parameter class $q \in \mathcal{E}$, as introduced in Equation (9). For each $q \in \mathcal{E}$, the orbit invariant $\boldsymbol{\beta}_q$ fixes the aggregate rescaled readout contributed by that class, while leaving open how this aggregate is distributed among neurons of its two opposite orientations. For each essential class $q \in \mathcal{E}$, define the orientation index sets

$$\mathcal{J}_q^+ := \{j \in \mathcal{J}_q \mid \overline{\mathbf{w}}_j \in \mathbb{R}_{>0}\overline{\mathbf{w}}_q\}, \qquad \mathcal{J}_q^- := \{j \in \mathcal{J}_q \mid \overline{\mathbf{w}}_j \in \mathbb{R}_{<0}\overline{\mathbf{w}}_q\}. \tag{19}$$

We consider readout allocations

$$\sum_{j\in\mathcal{J}_q^+} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = t_q\boldsymbol{\beta}_q, \qquad \sum_{j\in\mathcal{J}_q^-} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = (1-t_q)\boldsymbol{\beta}_q, \qquad t_q \in [0,1], \tag{20}$$

since every MWNP necessarily has such an allocation (Corollary F.2). For an allocation of this form, the contribution of essential class $q \in \mathcal{E}$ to the orbit residual $\mathbf{r}$ is

$$m\sum_{j\in\mathcal{J}_q} \mathbf{a}_j\overline{\mathbf{w}}_j^\top\overline{\mathbf{x}} = m\Big(\sum_{j\in\mathcal{J}_q^+} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j - \sum_{j\in\mathcal{J}_q^-} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j\Big)\overline{\mathbf{w}}_q^\top\overline{\mathbf{x}} = m(2t_q-1)\boldsymbol{\beta}_q\overline{\mathbf{w}}_q^\top\overline{\mathbf{x}}. \tag{21}$$

Define $\mathcal{T} \subseteq [0,1]^{|\mathcal{E}|}$ as the set of splits $\mathbf{t} = (t_q)_{q\in\mathcal{E}}$ for which these nonnegative allocations of the class aggregates $\boldsymbol{\beta}_q$ between orientations reproduce the entire orbit residual $\mathbf{r}$, that is,

$$\mathcal{T} := \Big\{\mathbf{t} \in [0,1]^{|\mathcal{E}|} \;\Big|\; m\sum_{q\in\mathcal{E}}(2t_q-1)\boldsymbol{\beta}_q\overline{\mathbf{w}}_q^\top\overline{\mathbf{x}} = \mathbf{r}(\mathbf{x}) \quad \text{for every } \mathbf{x} \in \mathbb{R}^{N_i}\Big\}, \tag{22}$$

and let $\mathcal{T}_{N_h} \subseteq \mathcal{T}$ denote the splits realizable using at most N_h neurons from essential classes.

Proposition 6.1 (Minimum-weight-norm representational geometry)**.** *Fix a symmetry orbit $\mathcal{O}$ with positively 1-homogeneous activation and essential parameter classes $\mathcal{E}$. If $\mathcal{T}_{N_h} \neq \emptyset$, the minimum of Ω_W over $\mathcal{O}_{N_h}$ is attained and equals $2\sum_{q\in\mathcal{E}}\|\boldsymbol{\beta}_q\|$. Moreover, the RSMs of all MWNPs in $\mathcal{O}_{N_h}$ are precisely*

$$\sum_{q\in\mathcal{E}}\|\boldsymbol{\beta}_q\|\big(t_q\mathbf{f}_q^+(\mathbf{f}_q^+)^\top + (1-t_q)\mathbf{f}_q^-(\mathbf{f}_q^-)^\top\big), \tag{23}$$

as $\mathbf{t}$ ranges over $\mathcal{T}_{N_h}$. (Proof in Appendix F.1.)

Under the feasibility condition $\mathcal{T}_{N_h} \neq \emptyset$, minimum-weight-norm selection therefore removes both degrees of freedom in representational geometry identified in Section 5.1. Auxiliary contributions to the RSM vanish, and the weight with which an essential class $q \in \mathcal{E}$ contributes to the RSM is fixed by the orbit invariant $\boldsymbol{\beta}_q$. Since $\boldsymbol{\beta}_q$ is the class's aggregate rescaled readout, the same orbit invariant that determines its contribution to the realized function also determines the weight of its contribution to the RSM, directly linking representational geometry to the network's computation.

The only remaining degree of freedom in Equation (23) is the split $\mathbf{t}$. For each $q \in \mathcal{E}$, the scalar t_q determines how the coefficient $\|\boldsymbol{\beta}_q\|$ is divided between the rank-one terms associated with the two feature vectors $\mathbf{f}_q^+$ and $\mathbf{f}_q^-$. Hence, MWNPs exhibit a unique RSM whenever the residual constraint in Equation (22) admits a unique split. This is guaranteed when $m \neq 0$ and the matrices $\boldsymbol{\beta}_q\overline{\mathbf{w}}_q^\top$, $q \in \mathcal{E}$, are linearly independent.

Corollary 6.2 (Minimum-weight-norm identifiability)**.** *In the setting of Proposition 6.1, suppose additionally that the matrices $\boldsymbol{\beta}_q \overline{\mathbf{w}}_q^\top$, $q \in \mathcal{E}$, are linearly independent and that $m \neq 0$. Then, all MWNPs in $\mathcal{O}$ of width at least N_h induce the same RSM.* (Proof in Appendix F.1.)

Although the set of representational geometries within a symmetry orbit can expand considerably with width (Proposition 5.2), minimum-weight-norm selection can single out a unique geometry that remains unchanged even as width increases. In particular, all six analytical ReLU solutions in Figure 1 satisfy the conditions of Corollary 6.2 (Example F.5). Within each corresponding symmetry orbit, all MWNPs therefore induce the same RSM across all widths $N_h \geq 2$, despite the substantial freedom permitted without minimum-norm selection. Thus, the very examples that expose the limitations of function alone in constraining representation also demonstrate how a principled selection rule can restore representational identifiability.

7 Discussion

It is common knowledge that function underdetermines the parameterization of artificial neural networks [19, 20, 22, 29, 30, 31, 32]. Building on work that makes this underdetermination precise by enumerating parameter symmetries in nonlinear networks [12, 24], we examine instead the underdetermination of *representational geometry* (Section 4). We characterize how parameter symmetries can reshape representational geometry while leaving function unchanged, and show that the range of similarity scores to a fixed reference systematically expands with overparameterization (Section 5), providing a theoretical account of dissociations observed empirically in prior work [17, 33, 34, 35, 36]. Finally, we demonstrate that the coupling between representation and function can be restored for certain implementations characterized by efficiency constraints (Section 6), mirroring a result from the linear regime [18]. Though our work is theoretical at present, it holds consequences for the affordances of neural representations; we comment on three.

Completeness. Şimşek et al. [12] establish a *complete* characterization of the global minima manifold for a teacher–student learning problem under specific assumptions on the activation function and the input distribution, showing that *all* zero-population-loss solutions lie in the orbit of the irreducible teacher under the symmetries they consider. Martinelli et al. [24] extend this framework to additional activation functions that admit further parameter symmetries, but do not establish that the resulting symmetry orbits exhaust the corresponding zero-loss sets. Further, both studies examine identifiability *within* a layer, and thus do not address degeneracies *across* layers, such as collapse of the representational geometry in an intermediate layer of a deep network [37, mechanism (iv)]. Accordingly, our results characterize representational variation generated by these parameter symmetries, without claiming that they exhaust the full fiber of a realized function.

Contravariance and task complexity. *Universal* representations have been observed across artificial neural networks trained on different tasks and modalities, and even across different architectures [38, 39, 40]. One proposed explanation is that the complexity of a function is *contravariant* to the "dispersion" (variability) of its implementations: harder problems admit fewer solutions [41]. This principle could explain why networks trained on increasingly broad or complex tasks converge toward universal representations [42, 43, 44]. Our results, however, demonstrate that contravariance need not hold for overparameterized implementations without additional implementation-level constraints. In a proportional regime, where model size grows alongside task complexity (as is arguably the case in practice [45]), the space of implementations need not narrow as the task becomes more complex. Contravariance, as a principle at the level of *function* rather than *implementation*, therefore cannot by itself account for representational universality.

Individual differences. Idiosyncrasies in representational geometry persist even among models of the same architecture [46, 47, 48, 49]. A central goal of cognitive computational neuroscience is nevertheless to isolate differences in neural activity that meaningfully distinguish individuals [50, 51], while recognizing that meaningfulness is necessarily context-dependent [52]. Our characterization of function-preserving variation in representational geometry in Section 5 provides a framework for distinguishing inter-individual differences that may be inconsequential from those that reflect genuine differences in computation. In particular, the normative selection rules of Section 6 can distinguish individuals along computationally relevant dimensions such as noise robustness and transfer performance [18, 53].

Acknowledgments and disclosure of funding

We thank Flavio Martinelli for helpful discussions that clarified aspects of the parameter symmetries cataloged in Martinelli et al. [24], and Felix A. Wichmann for valuable feedback on the manuscript. MT additionally thanks Felix A. Wichmann for his continued guidance and support throughout MT's doctoral studies.

MT was supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP INF, project number: 276693517. LB was supported by the Allen Institute. AMS was supported by a Schmidt Science Polymath Award, a Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z), and the Gatsby Charitable Foundation (GAT3850). EG was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2026-07018 and DGECR-2026-00191). AMS is a CIFAR Fellow in Learning in Machines & Brains, and EG is a CIFAR Azrieli Global Scholar in Learning in Machines & Brains and a Canada CIFAR AI Chair.

References

[1] M. D. Zeiler & R. Fergus. (2014). "Visualizing and understanding convolutional networks". In: *13th European Conference on Computer Vision*. Vol. 8689. Lecture Notes in Computer Science. Zurich, Switzerland: Springer, pp. 818–833.

[2] C. Olah et al. (2020). "Zoom in: An introduction to circuits". In: *Distill*.

[3] N. Saphra & S. Wiegreffe. (2024). "Mechanistic?" In: *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Miami, FL, USA: Association for Computational Linguistics, pp. 480–498.

[4] A. Mueller et al. (2026). "The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis". In: *Computational Linguistics* **52**(1), pp. 331–378.

[5] J. V. Haxby, A. C. Connolly & J. S. Guntupalli. (2014). "Decoding neural representational spaces using multivariate pattern analysis". In: *Annual Review of Neuroscience* **37**(1), pp. 435–456.

[6] N. Kriegeskorte, M. Mur & P. A. Bandettini. (2008). "Representational similarity analysis – connecting the branches of systems neuroscience". In: *Frontiers in Systems Neuroscience* **2**.

[7] D. L. K. Yamins & J. J. DiCarlo. (2016). "Using goal-driven deep learning models to understand sensory cortex". In: *Nature Neuroscience* **19**(3), pp. 356–365.

[8] K. Fakhar et al. (2024). "Downstream network transformations dissociate neural activity from causal functional contributions". In: *Scientific Reports* **14**, 2103.

[9] L. Albantakis et al. (2024). "The brain's best kept secret is its degenerate structure". In: *The Journal of Neuroscience* **44**(40), e1339242024.

[10] D. Kunin et al. (2021). "Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics". In: *9th International Conference on Learning Representations*. Virtual Event.

[11] R. Entezari, H. Sedghi, O. Saukh & B. Neyshabur. (2022). "The role of permutation invariance in linear mode connectivity of neural networks". In: *10th International Conference on Learning Representations*. Virtual Event.

[12] B. Şimşek et al. (2021). "Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. Virtual Event: PMLR, pp. 9722–9732.

[13] K. Hermann & A. K. Lampinen. (2020). "What shapes feature representations? Exploring datasets, architectures, and training". In: *Advances in Neural Information Processing Systems*. Vol. 33. Vancouver, Canada: Curran Associates, Inc., pp. 9995–10006.

[14] T. Flesch et al. (2022). "Orthogonal representations for robust context-dependent task performance in brains and neural networks". In: *Neuron* **110**(7), pp. 1258–1270.

[15] M. Farrell, S. Recanatesi & E. Shea-Brown. (2023). "From lazy to rich to exclusive task representations in neural networks and neural codes". In: *Current Opinion in Neurobiology* **83**, 102780.

[16] C.-N. Chou, H. Le, Y. Wang & S. Chung. (2025). "Feature learning beyond the lazy-rich dichotomy: Insights from representational geometry". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. Vancouver, Canada: PMLR, pp. 10700–10740.

[17] A. K. Lampinen, S. C. Y. Chan, Y. Li & K. Hermann. (2026). "Representation biases: Variance is not always a good proxy for importance". In: *eNeuro* **13**(3), ENEURO.0461-25.2026.

[18] L. Braun, E. Grant & A. M. Saxe. (2025). "Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks". In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. Vancouver, Canada: PMLR, pp. 5355–5382.

[19] R. Hecht-Nielsen. (1990). "On the algebraic structure of feedforward network weight spaces". In: *Advanced Neural Computers*. Ed. by R. Eckmiller. North-Holland, pp. 129–135. ISBN: 978-0-444-88400-8.

[20] H. J. Sussmann. (1992). "Uniqueness of the weights for minimal feedforward nets with a given input-output map". In: *Neural Networks* **5**(4), pp. 589–593.

[21] A. M. Chen, H.-m. Lu & R. Hecht-Nielsen. (1993). "On the geometry of feedforward neural network error surfaces". In: *Neural Computation* **5**(6), pp. 910–927.

[22] B. Neyshabur, R. R. Salakhutdinov & N. Srebro. (2015). "Path-SGD: Path-normalized optimization in deep neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 28. Montréal, Canada: Curran Associates, Inc., pp. 2422–2430.

[23] B. Zhao, R. Walters & R. Yu. (2026). "Symmetry in neural network parameter spaces". In: *Transactions on Machine Learning Research*.

[24] F. Martinelli, B. Şimşek, W. Gerstner & J. Brea. (2024). "Expand-and-Cluster: Parameter recovery of neural networks". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. Vienna, Austria: PMLR, pp. 34895–34919.

[25] S. Edelman. (1998). "Representation is representation of similarities". In: *Behavioral and Brain Sciences* **21**(4), pp. 449–467.

[26] N. Kriegeskorte & R. A. Kievit. (2013). "Representational geometry: Integrating cognition, computation, and the brain". In: *Trends in Cognitive Sciences* **17**(8), pp. 401–412.

[27] S. Kornblith, M. Norouzi, H. Lee & G. E. Hinton. (2019). "Similarity of neural network representations revisited". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. Long Beach, CA, USA: PMLR, pp. 3519–3529.

[28] A. H. Williams. (2024). "Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis". In: *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*. Vol. 285. Proceedings of Machine Learning Research. Vancouver, Canada: PMLR, pp. 10–23.

[29] V. Kůrková & P. C. Kainen. (1994). "Functionally equivalent feedforward neural networks". In: *Neural Computation* **6**(3), pp. 543–558.

[30] L. Dinh, R. Pascanu, S. Bengio & Y. Bengio. (2017). "Sharp minima can generalize for deep nets". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia: PMLR, pp. 1019–1028.

[31] D. M. Elbrächter, J. Berner & P. Grohs. (2019). "How degenerate is the parametrization of neural networks with the ReLU activation function?" In: *Advances in Neural Information Processing Systems*. Vol. 32. Vancouver, Canada: Curran Associates, Inc., pp. 7790–7801.

[32] M. Phuong & C. H. Lampert. (2020). "Functional vs. parametric equivalence of ReLU networks". In: *8th International Conference on Learning Representations*. Virtual Event.

[33] A. K. Lampinen, S. C. Y. Chan & K. Hermann. (2024). "Learned feature representations are biased by complexity, learning order, position, and more". In: *Transactions on Machine Learning Research*.

[34] N. Cloos et al. (2025). "Differentiable optimization of similarity scores between models and brains". In: *13th International Conference on Learning Representations*. Singapore, pp. 63438–63457.

[35] Y. Bo, A. Soni, S. Srivastava & M. Khosla. (2025). "Evaluating representational similarity measures from the lens of functional correspondence". In: *Proceedings of the 8th Annual Conference on Cognitive Computational Neuroscience*. Amsterdam, The Netherlands.

[36] M. Theiss, B. Fausten & F. A. Wichmann. (2026). "Representation can dissociate from function and behaviour in task-optimised vision networks". In: *9th Annual Conference on Cognitive Computational Neuroscience (Extended Abstract)*. New York, NY, USA.

[37] J. E. Grigsby, K. Lindsey & D. Rolnick. (2023). "Hidden symmetries of ReLU networks". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. Honolulu, HI, USA: PMLR, pp. 11734–11760.

[38] L. van Rossem & A. M. Saxe. (2024). "When representations align: Universality in representation learning dynamics". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. Vienna, Austria: PMLR, pp. 49098–49121.

[39] M. Huh, B. Cheung, T. Wang & P. Isola. (2024). "Position: The Platonic representation hypothesis". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. Vienna, Austria: PMLR, pp. 20617–20642.

[40] Z. Chen & M. F. Bonner. (2025). "Universal dimensions of visual representation". In: *Science Advances* **11**(27), eadw7697.

[41] R. Cao & D. L. K. Yamins. (2024). "Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle". In: *Cognitive Systems Research* **85**, 101200.

[42] Y. Li et al. (2016). "Convergent learning: Do different neural networks learn the same representations?" In: *4th International Conference on Learning Representations*. San Juan, Puerto Rico.

[43] Y. Bansal, P. Nakkiran & B. Barak. (2021). "Revisiting model stitching to compare neural representations". In: *Advances in Neural Information Processing Systems*. Vol. 34. Virtual Event: Curran Associates, Inc., pp. 225–236.

[44] C. Wolfram & A. Schein. (2025). "Layers at similar depths generate similar activations across LLM architectures". In: *2nd Conference on Language Modeling*. Montréal, Canada.

[45] Y. Bahri et al. (2024). "Explaining neural scaling laws". In: *Proceedings of the National Academy of Sciences* **121**(27), e2311878121.

[46] M. Schrimpf et al. (2020). "Integrative benchmarking to advance neurally mechanistic models of human intelligence". In: *Neuron* **108**(3), pp. 413–423.

[47] C. Conwell et al. (2024). "A large-scale examination of inductive biases shaping high-level visual representation in brains and machines". In: *Nature Communications* **15**(1), 9383.

[48] G. Tuckute, J. Feather, D. Boebinger & J. H. McDermott. (2023). "Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions". In: *PLOS Biology* **21**(12), e3002366.

[49] D. Linsley et al. (2023). "Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex". In: *Advances in Neural Information Processing Systems*. Vol. 36. New Orleans, LA, USA: Curran Associates, Inc., pp. 28873–28891.

[50] I. Thobani et al. (2025). "Model-brain comparison using inter-animal transforms". In: *Proceedings of the 8th Annual Conference on Cognitive Computational Neuroscience*. Amsterdam, The Netherlands.

[51] J. Feather, M. Khosla, N. A. R. Murty & A. Nayebi. (2025). *Brain-model evaluations need the NeuroAI Turing test*. arXiv preprint.

[52] B. Baker et al. (2026). "Use and usability: Concepts of representation in philosophy, neuroscience, cognitive science, and computer science". In: *Neurons, Behavior, Data Analysis, and Theory*.

[53] E. Holton et al. (2026). "Humans and neural networks show similar patterns of transfer and interference during continual learning". In: *Nature Human Behaviour* **10**(1), pp. 111–125.

[54] J. Bradbury et al. (2018). *JAX: Composable transformations of Python+NumPy programs*. Version 0.10.0.

[55] P. Kidger & C. Garcia. (2021). "Equinox: Neural networks in JAX via callable PyTrees and filtered transformations". In: *Differentiable Programming Workshop at the 35th Conference on Neural Information Processing Systems*. Virtual Event.

Appendix Contents

A Notation **16**

B Parameter symmetries in overparameterized nonlinear networks **20**

B.1 Generic reparameterization symmetries 20

B.1.1 Permutation symmetry 20

B.1.2 Positive-scaling symmetry 20

B.1.3 Sign-flip symmetry 20

B.2 Activation-independent overparameterization symmetries 21

B.2.1 Function-preserving realizations 21

B.2.2 Minimality and distinctness 22

B.3 Activation-dependent overparameterization symmetries 22

B.3.1 Even-linear and constant-odd activations 23

B.3.2 Aligned/opposite groups 23

B.3.3 Symmetries arising from even-linear activations 24

B.3.4 Symmetries arising from constant-odd activations 24

B.3.5 Choice of sign in aligned/opposite groups 24

B.3.6 Function-preserving realizations 25

B.3.7 Minimality and nondegeneracy 26

B.3.8 Nonuniqueness of irreducible parameterizations 27

B.4 Parameter classes, orbit invariants, and symmetry-equivalence 28

B.4.1 Parameter classes 29

B.4.2 Characterization of symmetry-equivalence 30

C Symmetry properties of common activation functions **33**

C.1 Even-linear activations 33

C.2 Constant-odd activations 37

C.3 Positively 1-homogeneous activations 39

C.4 Activations with neither symmetry property 39

D Feature transformations within symmetry orbits **41**

D.1 Duplication matrices 41

D.2 Feature-level characterization of parameter symmetries 43

D.3 Hidden-activation matrices within a symmetry orbit 45

D.4 Persistence of irreducible features 47

E Representational geometry and similarity within symmetry orbits **49**

E.1 RSM decomposition and positive reweighting 50

E.2 Strict growth until saturation 51

E.3 Function-independent similarity limits 55

E.4 Alignment with reference geometries 57

F Identifiability through minimum-norm selection **62**

F.1 Minimum weight norm for positively 1-homogeneous activations 62

F.2 Minimum representation norm for positively 1-homogeneous activations 68

G Numerical and analytical characterization of ReLU networks solving the XOR task **74**

G.1 Setup 75

G.2 Gradient-descent sweep and solution clustering 75

G.3 Closed-form characterization of the six dominant solution types 77

G.4 Approaching zero BCE loss by parameter scaling 79

A Notation

This appendix collects the notation used throughout the paper, organized by topic and sorted alphabetically within each block. We adopt the following standard conventions:

- lowercase letters (e.g., b_j, c, m) denote scalars,
- bold lowercase letters (e.g., $\mathbf{w}_j$, $\mathbf{a}_j$, $\mathbf{h}^\mu$) denote vectors,
- and bold uppercase letters (e.g., $\mathbf{W}$, $\mathbf{A}$, $\mathbf{H}$) denote matrices.

When corresponding quantities are defined for both an arbitrary parameterization and an irreducible parameterization, the latter are marked by a superscript star; for example, $\mathbf{w}_j^\star$, $b_j^\star$, $\mathbf{a}_j^\star$ (irreducible) vs. $\mathbf{w}_j$, b_j, $\mathbf{a}_j$ (arbitrary).

Indices and dimensions

j	Index enumerating neurons
K	Number of rows of $\mathbf{U}$
μ	Index enumerating inputs
N_h, $N_h^\star$	Width of a parameterization, and minimum width within a symmetry orbit, respectively
N_i	Input dimension
N_o	Output dimension
P	Number of inputs

Constants, norms, inner products, and operators

$\mathbf{0}$, $\mathbf{1}$	Zero vector or matrix, all-ones vector
$\lvert\cdot\rvert$	Absolute value, or set cardinality
$\mathrm{diag}(\cdot)$	Diagonal matrix from a vector, or extraction of the diagonal of a matrix
$\langle \mathbf{M}, \mathbf{N} \rangle_F$	Frobenius inner product $\mathrm{tr}(\mathbf{M}^\top \mathbf{N})$
$\lVert\cdot\rVert$	Euclidean norm
$\lVert\cdot\rVert_F$	Frobenius norm
$\rho(\mathbf{M}, \mathbf{N})$	Pearson correlation between strict upper-triangular entries of two RSMs

Network parameters

$\mathbf{a}_j$, $\mathbf{a}_j^\star$	Readout weights of a single neuron
$\mathbf{a}^\pm$, $\mathbf{a}^\mp$	Sum of and difference between aggregate readouts of aligned and opposite subgroups
$\mathbf{a}_\mathcal{J}^\pm$, $\mathbf{a}_\mathcal{J}^\mp$	Corresponding sum and difference restricted to a subset $\mathcal{J}$ of neurons
$\mathbf{A}$, $\mathbf{A}^\star$	Readout-weight matrix
$\mathbf{b}$, $\mathbf{b}^\star$	Hidden-bias vector
b_j, $b_j^\star$	Hidden bias of a single neuron
$f_{\boldsymbol{\theta}}$	Function realized by the network with parameterization $\boldsymbol{\theta}$
$\boldsymbol{\theta}$, $\boldsymbol{\theta}^\star$	Network parameterization consisting of the nonlinearity σ, incoming weights $\mathbf{W}$, hidden biases $\mathbf{b}$, and readout weights $\mathbf{A}$

$\mathbf{w}_j$, $\mathbf{w}_j^\star$	Incoming-weight vector of a single neuron
$\overline{\mathbf{w}}_j$, $\overline{\mathbf{w}}_j^\star$	Incoming-parameter vector $(\mathbf{w}_j^\top, b_j)^\top \in \mathbb{R}^{N_i+1}$ of a single neuron
$\mathbf{W}$, $\mathbf{W}^\star$	Incoming-weight matrix

Inputs and hidden activations

$\mathbf{f}_\ell^\top$	ℓth row of $\mathbf{F}_{\mathcal{E},\mathcal{I}}$
$(\mathbf{f}_q^+)^\top$, $(\mathbf{f}_q^-)^\top$	Rows of $\mathbf{F}_\mathcal{E}$ corresponding to the two orientations $\pm\overline{\mathbf{w}}_q$ of essential parameter class q
$\mathbf{F}_\mathcal{E}$, $\mathbf{F}_{\mathcal{E},\mathcal{I}}$	Feature matrices for both orientations of every essential parameter class and for the subset of represented orientations indexed by $\mathcal{I}$
$\mathbf{h}^\mu$	Hidden activation elicited by input $\mathbf{x}^\mu$
$\mathbf{H}$, $\mathbf{H}^\star$	Hidden-activation matrix
$\mathcal{I}$	Index set selecting the rows of $\mathbf{F}_\mathcal{E}$ represented in a parameterization
$\mathbf{u}_k^\top$	kth row of $\mathbf{U}$
$\mathbf{U}$	Matrix collecting the pairwise distinct additional feature rows contributed by nonessential parameter classes and constant neurons
$\mathbf{v}_j^\top$, $\mathbf{v}_j^{\star\top}$	jth feature row of $\mathbf{H}$ or $\mathbf{H}^\star$, respectively
$\mathbf{x}^\mu$	μth input vector
$\overline{\mathbf{x}}$	Augmented input vector $(\mathbf{x}^\top, 1)^\top$
$\mathbf{X}$	Input matrix $[\mathbf{x}^1, \dots, \mathbf{x}^P]$ collecting all P inputs as columns
$\overline{\mathbf{X}}$	Augmented input matrix $[\overline{\mathbf{x}}^1, \dots, \overline{\mathbf{x}}^P] = [\mathbf{X}^\top, \mathbf{1}]^\top$
z, z_j	Pre-activation $\overline{\mathbf{w}}^\top \overline{\mathbf{x}}$ of a neuron

Activation functions

c	Constant value of the even component of a constant-odd activation
δ	Coefficient of the absolute-value component of a positively 1-homogeneous activation
$e(z)$, $o(z)$	Even and odd components of a function
λ_-, λ_+	Slopes of a positively 1-homogeneous activation on the negative and positive half-lines, respectively
m	Slope of the linear component of an even-linear activation
$\psi(z)$	ReLU activation $\psi(z) = \max(0, z)$
$\sigma\colon \mathbb{R} \to \mathbb{R}$	Nonaffine scalar activation function

Symmetry groups and orbits

$\mathcal{D}$	Index set of a duplicate-neuron group
$\mathcal{K}$	Index set of an aligned/opposite group
$\mathcal{N}^+$, $\mathcal{N}^-$	Index sets of aligned and opposite subgroups of an aligned/opposite group
$\mathcal{O}(\boldsymbol{\theta})$, $\mathcal{O}_{N_h}(\boldsymbol{\theta})$	Symmetry orbit of $\boldsymbol{\theta}$ (at width N_h)

$\boldsymbol{\theta} \sim \boldsymbol{\xi}$	Symmetry-equivalence of network parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$
$\mathcal{Z}$	Index set of a zero-neuron group

Parameter classes and orbit invariants

$\boldsymbol{\beta}_q(\boldsymbol{\theta})$, $\boldsymbol{\beta}_q$	Class aggregate of parameter class q in $\boldsymbol{\theta}$ (the aggregate rescaled readout for positively 1-homogeneous activations), and its common value on a fixed symmetry orbit
$\mathcal{E}$	Orbit-invariant set of essential parameter classes, characterized by $\boldsymbol{\beta}_q \neq \mathbf{0}$
$\mathcal{J}_q$, $\mathcal{J}_0$	Index sets of the neurons in parameter class q and of constant neurons, respectively
$\mathcal{J}_q^+$, $\mathcal{J}_q^-$	Neurons of $\mathcal{J}_q$ whose incoming parameters $\overline{\mathbf{w}}_j$ are positive and negative multiples of $\overline{\mathbf{w}}_q$, respectively
$\phi_q(\mathbf{x})$	Nonlinear feature associated with parameter class q
q	Activation-dependent parameter class of a nonconstant neuron
$\mathcal{Q}(\boldsymbol{\theta})$	Set of parameter classes represented by the nonconstant neurons of $\boldsymbol{\theta}$
$\mathbf{r}_{\boldsymbol{\theta}}$, $\mathbf{r}$	Global residual function of $\boldsymbol{\theta}$, and its common value on a fixed orbit
$[\overline{\mathbf{w}}]$	Parameter class of the incoming-parameter vector $\overline{\mathbf{w}}$
$\overline{\mathbf{w}}_q$	Chosen incoming-parameter representative of class q, normalized to unit norm in the positively 1-homogeneous case, and with a positive first nonzero incoming-weight entry in the constant-odd case

Primitive feature transformations

$\boldsymbol{\alpha}$	Vector of nonzero scaling factors
$\boldsymbol{\alpha}^2$	Hadamard square of $\boldsymbol{\alpha}$
$\mathcal{A}_{\mathbf{U}}$, $\mathcal{D}_{\boldsymbol{\nu}}$, $\mathcal{S}_{\boldsymbol{\alpha}}$	Feature addition, duplication, and scaling, respectively
$\mathbf{D}_{\boldsymbol{\nu}}$	Binary duplication matrix that repeats the ith pre-duplication feature row exactly ν_i times
$\boldsymbol{\gamma} = \mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2$	Effective weights of the rank-one contributions to the RSM induced by duplication and scaling
$\boldsymbol{\nu}$	Duplication pattern, with $\nu_i \in \mathbb{N}_{>0}$ denoting the number of copies of the ith pre-duplication feature row

Representational geometry and similarity

cl	Closure with respect to the Frobenius norm
$\mathrm{cone}(\cdot)$, $\overline{\mathrm{cone}}(\cdot)$	Conic hull operator and its closure, respectively
$\mathcal{C}_{\sigma,\mathbf{X}}$	Closed conic hull of $\mathcal{G}_{\sigma,\mathbf{X}}$
d	Dimension of the span of $\mathcal{C}_{\sigma,\mathbf{X}}$
$\Delta_{N_h}(\mathbf{N})$, $\Delta_\infty(\mathbf{N})$	Spread of attainable similarity scores relative to $\mathbf{N}$ at width N_h and across all widths, respectively
$\mathcal{G}_{\sigma,\mathbf{X}}$	Set of projected rank-one contributions from all single-neuron features realizable with activation σ on $\mathbf{X}$

$\mathcal{H}$	Subspace of hollow symmetric matrices with zero off-diagonal mean
$\mathbf{M}_{\boldsymbol{\theta}}$	RSM induced by parameterization $\boldsymbol{\theta}$
$\mathbf{M}^{\circ}$	Normalized projection of a symmetric matrix $\mathbf{M}$ onto $\mathcal{H}$
$\mathbf{N}$	Fixed reference RSM used for representational comparison
$\Pi_{\mathcal{H}}$	Orthogonal projection onto $\mathcal{H}$
$\mathcal{R}_{N_h}, \mathcal{R}$	Sets of realizable RSMs within the symmetry orbit at width N_h and across all widths, respectively
$\mathcal{S}_{N_h}(\mathbf{N}), \mathcal{S}(\mathbf{N})$	Sets of similarity scores to reference $\mathbf{N}$ attainable within the symmetry orbit at width N_h and across all widths, respectively

Readout allocation and norm minimization

g_q	Minimum of g_q^+ and g_q^-
g_q^+, g_q^-	Norms of the two features $\mathbf{f}_q^+$ and $\mathbf{f}_q^-$, respectively, of essential parameter class q
$\kappa(\mathbf{t})$	Minimum number of essential-class neurons required to realize the split $\mathbf{t}$
$\Omega_H(\boldsymbol{\theta})$	Representation-norm objective $\|\mathbf{H}\|_F^2 + \|\mathbf{A}\|_F^2$
$\Omega_W(\boldsymbol{\theta})$	Weight-norm objective $\|\mathbf{W}\|_F^2 + \|\mathbf{b}\|^2 + \|\mathbf{A}\|_F^2$
$t_q \in [0,1]$	Fraction of $\boldsymbol{\beta}_q$ allocated to orientation $\overline{\mathbf{w}}_q$, with the remaining fraction $1 - t_q$ allocated to $-\overline{\mathbf{w}}_q$
$\mathbf{t} = (t_q)_{q \in \mathcal{E}}$	Vector of readout splits across essential parameter classes
$\mathcal{T}$	Set of residual-compatible splits, i.e., splits that reproduce the orbit residual $\mathbf{r}$
$\mathcal{T}_{N_h}$	Width-feasible subset of $\mathcal{T}$, comprising the splits satisfying $\kappa(\mathbf{t}) \leq N_h$
$\mathcal{T}^H, \mathcal{T}_{N_h}^H$	Splits in $\mathcal{T}$ and $\mathcal{T}_{N_h}$, respectively, that use only orientations of minimal feature norm g_q

B Parameter symmetries in overparameterized nonlinear networks

This appendix provides a formal treatment of the function-preserving parameter symmetries in overparameterized nonlinear networks that are summarized in the main text. We distinguish three classes of symmetries:

- generic reparameterization symmetries (Appendix B.1),
- symmetries induced by overparameterization that are independent of the activation function (Appendix B.2),
- and symmetries induced by overparameterization that depend on algebraic symmetries of the activation function (Appendix B.3).

After introducing all parameter symmetries considered in this work, we establish a necessary and sufficient characterization of when two parameterizations are symmetry-equivalent (Appendix B.4).

B.1 Generic reparameterization symmetries

We briefly revisit three well-known function-preserving symmetries of neural networks: the permutation symmetry (Appendix B.1.1), the positive-scaling symmetry (Appendix B.1.2), and the sign-flip symmetry (Appendix B.1.3).

B.1.1 Permutation symmetry

For a hidden layer of width N_h, the ordering of the neurons within that layer is arbitrary. Let $\mathfrak{S}_{N_h}$ denote the *symmetric group* of the set of integers $\{1, \dots, N_h\}$, and let $\pi \in \mathfrak{S}_{N_h}$ be a permutation. Reordering the parameters of the layer according to

$$(\mathbf{w}_j, b_j, \mathbf{a}_j)_{j=1}^{N_h} \mapsto (\mathbf{w}_{\pi(j)}, b_{\pi(j)}, \mathbf{a}_{\pi(j)})_{j=1}^{N_h} \tag{24}$$

does not change the function computed by the layer, since

$$\sum_{j=1}^{N_h} \mathbf{a}_{\pi(j)}\, \sigma(\mathbf{w}_{\pi(j)}^\top \mathbf{x} + b_{\pi(j)}) = \sum_{j=1}^{N_h} \mathbf{a}_j\, \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j), \qquad \mathbf{x} \in \mathbb{R}^{N_i}. \tag{25}$$

This invariance is commonly referred to as the *permutation symmetry*.

B.1.2 Positive-scaling symmetry

Suppose the activation function $\sigma \colon \mathbb{R} \to \mathbb{R}$ is positively homogeneous of degree 1, i.e.,

$$\sigma(\alpha z) = \alpha\, \sigma(z), \qquad \alpha > 0. \tag{26}$$

Rescaling the parameters of a single neuron via

$$(\mathbf{w}_j, b_j, \mathbf{a}_j) \mapsto (\alpha \mathbf{w}_j, \alpha b_j, \alpha^{-1} \mathbf{a}_j), \qquad \alpha > 0, \tag{27}$$

does not change the function computed by that neuron since, for $\mathbf{x} \in \mathbb{R}^{N_i}$,

$$\alpha^{-1} \mathbf{a}_j\, \sigma(\alpha \mathbf{w}_j^\top \mathbf{x} + \alpha b_j) = (\alpha^{-1}\alpha) \mathbf{a}_j\, \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j) = \mathbf{a}_j\, \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j). \tag{28}$$

This invariance is commonly referred to as the *positive-scaling symmetry*.

B.1.3 Sign-flip symmetry

Suppose the activation function $\sigma \colon \mathbb{R} \to \mathbb{R}$ is odd, i.e., $\sigma(-z) = -\sigma(z)$. Flipping the sign of both the incoming parameters $(\mathbf{w}_j, b_j)$ and the readout weights $\mathbf{a}_j$, i.e.,

$$(\mathbf{w}_j, b_j, \mathbf{a}_j) \mapsto (-\mathbf{w}_j, -b_j, -\mathbf{a}_j), \tag{29}$$

does not change the function computed by that neuron since, for $\mathbf{x} \in \mathbb{R}^{N_i}$,

$$-\mathbf{a}_j\, \sigma(-\mathbf{w}_j^\top \mathbf{x} - b_j) = -\mathbf{a}_j\, \sigma(-(\mathbf{w}_j^\top \mathbf{x} + b_j)) = \mathbf{a}_j\, \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j). \tag{30}$$

Similarly, if σ is even, i.e., $\sigma(-z) = \sigma(z)$, flipping the sign of the incoming parameters,

$$(\mathbf{w}_j, b_j, \mathbf{a}_j) \mapsto (-\mathbf{w}_j, -b_j, \mathbf{a}_j), \tag{31}$$

also leaves the realized function invariant. We refer to this as the *sign-flip symmetry*.

B.2 Activation-independent overparameterization symmetries

In contrast to the generic function-preserving reparameterization symmetries discussed in Appendix B.1, the symmetries considered in this and the subsequent subsections are induced by overparameterization: they rely on the presence of "redundant" neurons. We first review overparameterization symmetries that do *not* depend on the choice of activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$. In a two-layer, bias-free setting, Şimşek et al. [12] show that these symmetries generate affine subspaces of equivalent networks (an *expansion manifold*) in a teacher–student setup.

This subsection first introduces a nondegeneracy condition on readout weights that rules out decomposable symmetry groups, before defining the three activation-independent symmetry classes: zero-neuron groups, duplicate-neuron groups, and constant neurons. We then discuss how these symmetry classes give rise to function-preserving realizations (Appendix B.2.1), and demonstrate that the symmetry classes are minimal and mutually distinct (Appendix B.2.2).

Compared with prior formulations [12, 24], we impose the following nondegeneracy condition to exclude decomposable cases.

Definition B.1 (Subset-nonzero)**.** Let $\mathcal{I} \subseteq \{1, \dots, N_h\}$ be a finite index set. We say that the collection of readout weights $\{\mathbf{a}_i\}_{i\in\mathcal{I}}$ is *subset-nonzero* if, for every nonempty proper subset $\mathcal{J} \subsetneq \mathcal{I}$,

$$\sum_{j\in\mathcal{J}} \mathbf{a}_j \neq \mathbf{0}. \tag{32}$$

We now formalize the three activation-independent overparameterization symmetry classes considered in this subsection.

Definition B.2 (Zero-neuron group)**.** Fix $(\mathbf{w}, b) \in \mathbb{R}^{N_i} \times \mathbb{R}$ with $\mathbf{w} \neq \mathbf{0}$, and let $\mathcal{Z} \neq \emptyset$ index a set of hidden neurons satisfying $\mathbf{w}_z = \mathbf{w}$ and $b_z = b$ for all $z \in \mathcal{Z}$, with readout weights $\mathbf{a}_z$. We call the neurons indexed by $\mathcal{Z}$ a *zero-neuron group* if

$$\sum_{z\in\mathcal{Z}} \mathbf{a}_z = \mathbf{0} \qquad \text{and} \qquad \{\mathbf{a}_z\}_{z\in\mathcal{Z}} \text{ is subset-nonzero.} \tag{33}$$

Definition B.3 (Duplicate-neuron group)**.** Fix $(\mathbf{w}^\star, b^\star, \mathbf{a}^\star) \in \mathbb{R}^{N_i} \times \mathbb{R} \times \mathbb{R}^{N_o}$ with $\mathbf{w}^\star \neq \mathbf{0}$ and $\mathbf{a}^\star \neq \mathbf{0}$, and let $\mathcal{D}$ index a set of at least two hidden neurons satisfying $\mathbf{w}_d = \mathbf{w}^\star$ and $b_d = b^\star$ for all $d \in \mathcal{D}$, with readout weights $\mathbf{a}_d$. We call the neurons indexed by $\mathcal{D}$ a *duplicate-neuron group* if

$$\sum_{d\in\mathcal{D}} \mathbf{a}_d = \mathbf{a}^\star \qquad \text{and} \qquad \{\mathbf{a}_d\}_{d\in\mathcal{D}} \text{ is subset-nonzero.} \tag{34}$$

Definition B.4 (Constant neuron)**.** A *constant neuron* is a hidden neuron with vanishing incoming weights $\mathbf{w} = \mathbf{0}$.

B.2.1 Function-preserving realizations

Zero-neuron groups. Adding or removing a zero-neuron group to or from a parameterization leaves the realized function unchanged. Indeed,

$$\sum_{z\in\mathcal{Z}} \mathbf{a}_z\, \sigma(\mathbf{w}_z^\top \mathbf{x} + b_z) = \sigma(\mathbf{w}^\top \mathbf{x} + b) \underbrace{\sum_{z\in\mathcal{Z}} \mathbf{a}_z}_{=\mathbf{0}} = \mathbf{0}. \tag{35}$$

Duplicate-neuron groups. The same holds when a neuron with parameters $(\mathbf{w}^\star, b^\star, \mathbf{a}^\star)$ is replaced by a duplicate-neuron group, since

$$\sum_{d\in\mathcal{D}} \mathbf{a}_d\, \sigma(\mathbf{w}_d^\top \mathbf{x} + b_d) = \sigma({\mathbf{w}^\star}^\top \mathbf{x} + b^\star) \underbrace{\sum_{d\in\mathcal{D}} \mathbf{a}_d}_{=\mathbf{a}^\star} = \mathbf{a}^\star\, \sigma({\mathbf{w}^\star}^\top \mathbf{x} + b^\star). \tag{36}$$

Merging copies of a duplicate-neuron group into a single neuron leaves the realized function invariant as well.

Constant neurons. A constant neuron, by contrast, contributes an input-independent offset:

$$\mathbf{a}\,\sigma(\mathbf{w}^\top \mathbf{x} + b) = \mathbf{a}\,\sigma(b). \tag{37}$$

Thus, for the realized function to remain unchanged, the constant contributions of all such neurons must cancel. If σ is constant-odd (Definition B.6), this offset may also be compensated by constant-neuron groups (Definition B.11) or constant-duplicate-neuron groups (Definition B.12).

B.2.2 Minimality and distinctness

Requiring that the shared incoming-weight vector be nonzero for zero-neuron and duplicate-neuron groups, and incorporating the subset-nonzero condition from Definition B.1 into their definitions, guarantees that the resulting taxonomy of activation-independent symmetries is *minimal* in the sense that groups do not decompose into smaller subgroups.

Zero-neuron groups. Let $\mathcal{Z}$ index a zero-neuron group. Then, the subset-nonzero condition guarantees that any nonempty proper subset $\mathcal{J} \subsetneq \mathcal{Z}$ satisfies $\sum_{j\in\mathcal{J}} \mathbf{a}_j \neq \mathbf{0}$. Hence, a zero-neuron group can never contain a proper zero-neuron subgroup. In other words, Definition B.2 singles out the smallest groups of neurons that could be removed from a parameterization without altering its realized function.

Duplicate-neuron groups. Let $\mathcal{D}$ index a duplicate-neuron group with shared parameters $(\mathbf{w}^\star, b^\star)$ and aggregate readout weights $\sum_{d\in\mathcal{D}} \mathbf{a}_d = \mathbf{a}^\star \neq \mathbf{0}$. Then, there exists no proper duplicate-neuron subgroup *with respect to* $\mathbf{a}^\star$,[3] i.e., there does not exist a subset of neurons $\mathcal{J} \subsetneq \mathcal{D}$ such that $\sum_{j\in\mathcal{J}} \mathbf{a}_j = \mathbf{a}^\star$. To see this, assume the opposite. The identity $\sum_{j\in\mathcal{J}} \mathbf{a}_j = \mathbf{a}^\star$ implies

$$\sum_{d\in\mathcal{D}\setminus\mathcal{J}} \mathbf{a}_d = \sum_{d\in\mathcal{D}} \mathbf{a}_d - \sum_{j\in\mathcal{J}} \mathbf{a}_j = \mathbf{a}^\star - \mathbf{a}^\star = \mathbf{0}, \tag{38}$$

violating the subset-nonzero condition. As for zero-neuron groups, the subset-nonzero condition also guarantees that a duplicate-neuron group can never contain a zero-neuron subgroup.

Constant neurons. Since constant neurons are individual neurons, rather than groups of neurons, they trivially do not admit a decomposition into subgroups.

Definitions B.2 to B.4 not only satisfy this minimality property, but also yield a proper taxonomy of activation-independent overparameterization symmetries: the three classes are mutually *distinct*. Zero-neuron groups and duplicate-neuron groups require the shared incoming weights to be nonzero, which guarantees that both are distinct from constant neurons, since the latter are defined by having vanishing incoming weights $\mathbf{w} = \mathbf{0}$. Similarly, a zero-neuron group never qualifies as a duplicate-neuron group because the latter requires the aggregate readout $\mathbf{a}^\star$ to be nonzero. Hence, Definitions B.2 to B.4 are all mutually exclusive.

B.3 Activation-dependent overparameterization symmetries

The three activation-*independent* symmetries in Appendix B.2 arise from overparameterization alone, regardless of the activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$. Additional symmetry groups, composed of "aligned" and "opposite" subgroups of neurons whose incoming weights and hidden biases agree up to sign flips, emerge when σ has algebraic structure relating $\sigma(z)$ and $\sigma(-z)$. Formalizing these symmetries requires two ingredients: we first identify two classes of activation functions that exhibit the relevant algebraic structure (Appendix B.3.1), and then formalize the notion of a group of neurons splitting into aligned and opposite subgroups (Appendix B.3.2). We then review the corresponding symmetry groups for even-linear activations (Appendix B.3.3) and constant-odd activations (Appendix B.3.4), before discussing the choice of sign in aligned/opposite subgroups (Appendix B.3.5), their function-preserving realizations (Appendix B.3.6), and the degenerate cases excluded by the nondegeneracy requirements placed on aligned/opposite subgroups (Appendix B.3.7). Finally, we demonstrate that, when σ admits activation-dependent overparameterization symmetries, irreducible parameterizations can differ by more than generic reparameterization symmetries alone (Appendix B.3.8).

[3] In contrast to zero-neuron groups, for duplicate-neuron groups the notion of minimality is only sensible with respect to a fixed reference readout $\mathbf{a}^\star$, since *any* subset $\mathcal{J} \subsetneq \mathcal{D}$ indexing at least two neurons again defines a duplicate-neuron group with respect to $\mathbf{a}^\star_{\mathcal{J}} \coloneqq \sum_{j\in\mathcal{J}} \mathbf{a}_j$.

B.3.1 Even-linear and constant-odd activations

Recall that any function $\sigma\colon \mathbb{R} \to \mathbb{R}$ can be uniquely decomposed into even and odd components $\sigma(z) = e(z) + o(z)$, where

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2}, \qquad o(z) = \frac{\sigma(z) - \sigma(-z)}{2}, \tag{39}$$

satisfying $e(-z) = e(z)$ and $o(-z) = -o(z)$. Following Martinelli et al. [24], we distinguish two classes of activations by the structure of their even and odd components.

Definition B.5 (Even-linear activations)**.** An activation function is called *even-linear* if its odd component is linear: $\sigma(z) = e(z) + mz$ for $m \in \mathbb{R}$.

Definition B.6 (Constant-odd activations)**.** An activation function is called *constant-odd* if its even component is constant: $\sigma(z) = c + o(z)$ for $c \in \mathbb{R}$.

These two classes are not mutually exclusive: an activation is both even-linear and constant-odd if and only if it is affine, $\sigma(z) = mz + c$. As established in Section 2, we exclude affine activations throughout so that the two classes are mutually exclusive in our setting.

B.3.2 Aligned/opposite groups

Here, we formalize subgroups of hidden neurons that share incoming weights and hidden biases up to a sign flip as *aligned/opposite groups*. These underlie the activation-dependent symmetry groups of Martinelli et al. [24], which we revisit and refine by imposing additional nondegeneracy conditions on the aligned and opposite subgroups. Much like the subset-nonzero condition for activation-independent symmetries, these conditions prevent symmetry groups from decomposing into smaller ones. Without them, activation-*dependent* symmetry groups can be broken up into smaller activation-*dependent* symmetry groups or even collapse into disjoint activation-*independent* symmetry groups. A detailed discussion of these degenerate cases is provided in Appendix B.3.7.

Definition B.7 (Aligned/opposite group)**.** Fix $(\mathbf{w}, b) \in \mathbb{R}^{N_i} \times \mathbb{R}$ with $\mathbf{w} \neq \mathbf{0}$, and let $\mathcal{K}$ index a set of hidden neurons with parameters $(\mathbf{w}_k, b_k, \mathbf{a}_k)$ for $k \in \mathcal{K}$. Let $\mathcal{N}^+, \mathcal{N}^- \subsetneq \mathcal{K}$ be nonempty, disjoint index sets such that $\mathcal{K} = \mathcal{N}^+ \sqcup \mathcal{N}^-$. The group of neurons indexed by $\mathcal{K}$ is called an *aligned/opposite group* with respect to $(\mathbf{w}, b)$ if it splits into an aligned subgroup $\mathcal{N}^+$ and an opposite subgroup $\mathcal{N}^-$ such that

$$(\mathbf{w}_k, b_k) = (\mathbf{w}, b), \quad k \in \mathcal{N}^+, \qquad (\mathbf{w}_k, b_k) = (-\mathbf{w}, -b), \quad k \in \mathcal{N}^-. \tag{40}$$

Note that the splitting is defined only up to a global sign flip: the same set of neurons splits into an aligned subgroup $\mathcal{N}^-$ and an opposite subgroup $\mathcal{N}^+$ with respect to parameters $(-\mathbf{w}, -b)$. In Appendix B.3.5, we show that, for some of the symmetry groups to be introduced in Appendices B.3.3 and B.3.4, the labels "aligned" and "opposite" are indeed interchangeable, whereas for others they acquire a semantic meaning that removes this sign ambiguity altogether.

Next, we introduce two nondegeneracy conditions that yield *minimal* activation-*dependent* symmetry groups, paralleling the role of the subset-nonzero condition in ensuring minimality for activation-*independent* symmetry groups (Definition B.1).

Definition B.8 (Nondegeneracy of aligned/opposite groups)**.** Let $\mathcal{K}$ index an aligned/opposite group with subgroups indexed by $\mathcal{N}^+$ and $\mathcal{N}^-$. For every nonempty subset $\mathcal{J} \subseteq \mathcal{K}$, we define

$$\mathbf{a}_{\mathcal{J}}^{\pm} \coloneqq \sum_{k \in \mathcal{J} \cap \mathcal{N}^+} \mathbf{a}_k + \sum_{k \in \mathcal{J} \cap \mathcal{N}^-} \mathbf{a}_k, \qquad \mathbf{a}_{\mathcal{J}}^{\mp} \coloneqq \sum_{k \in \mathcal{J} \cap \mathcal{N}^+} \mathbf{a}_k - \sum_{k \in \mathcal{J} \cap \mathcal{N}^-} \mathbf{a}_k. \tag{41}$$

We write

$$\mathbf{a}^{\pm} \coloneqq \mathbf{a}_{\mathcal{K}}^{\pm} = \sum_{k \in \mathcal{N}^+} \mathbf{a}_k + \sum_{k \in \mathcal{N}^-} \mathbf{a}_k, \qquad \mathbf{a}^{\mp} \coloneqq \mathbf{a}_{\mathcal{K}}^{\mp} = \sum_{k \in \mathcal{N}^+} \mathbf{a}_k - \sum_{k \in \mathcal{N}^-} \mathbf{a}_k \tag{42}$$

for the aggregate weights of the full group. An aligned/opposite group is called *sum-nondegenerate* if $\mathbf{a}_{\mathcal{J}}^{\pm} \neq \mathbf{0}$ for every nonempty proper subset $\mathcal{J} \subsetneq \mathcal{K}$, and *difference-nondegenerate* if $\mathbf{a}_{\mathcal{J}}^{\mp} \neq \mathbf{0}$ for every such $\mathcal{J}$.

Both notions are invariant under the global sign flip described above: exchanging $\mathcal{N}^+$ and $\mathcal{N}^-$ leaves $\mathbf{a}_{\mathcal{J}}^{\pm}$ unchanged and negates $\mathbf{a}_{\mathcal{J}}^{\mp}$.

B.3.3 Symmetries arising from even-linear activations

Assume $\sigma(z) = e(z) + mz$ is even-linear in the sense of Definition B.5. For an aligned/opposite group with parameters $(\mathbf{w}, b)$, letting $z_k \coloneqq \mathbf{w}_k^\top \mathbf{x} + b_k$ and $z \coloneqq \mathbf{w}^\top \mathbf{x} + b$, we have $z_k = z$ for $k \in \mathcal{N}^+$ and $z_k = -z$ for $k \in \mathcal{N}^-$. Thus, the combined contribution of the aligned/opposite group to the layer output is

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \sum_k \mathbf{a}_k \, e(z_k) + \sum_k \mathbf{a}_k \, m z_k = \mathbf{a}^{\pm} e(z) + \mathbf{a}^{\mp} m z. \tag{43}$$

Hence, the aggregate readout weights $\mathbf{a}^{\pm}$ and $\mathbf{a}^{\mp}$ control the even and linear components of the group's contribution, respectively. In particular, an aligned/opposite group with $\mathbf{a}^{\pm} = \mathbf{0}$ generates a linear contribution, while an aligned/opposite group satisfying $\mathbf{a}^{\pm} = \mathbf{a}^\star$ reproduces the contribution of a single reference neuron with parameters $(\mathbf{w}, b, \mathbf{a}^\star)$ up to a residual linear contribution. This motivates the following two symmetry groups, which correspond to the "even + linear" symmetries of Martinelli et al. [24].

Definition B.9 (Linear-neuron group)**.** Suppose $\sigma(z) = e(z) + mz$ is even-linear. A *linear-neuron group* is a sum-nondegenerate aligned/opposite group with parameters $(\mathbf{w}, b)$ such that $\mathbf{a}^{\pm} = \mathbf{0}$.

Definition B.10 (Linear-duplicate-neuron group)**.** Suppose $\sigma(z) = e(z) + mz$ is even-linear, and fix parameters $(\mathbf{w}^\star, b^\star, \mathbf{a}^\star)$ with $\mathbf{a}^\star \neq \mathbf{0}$. A *linear-duplicate-neuron group* is a sum-nondegenerate aligned/opposite group with parameters $(\mathbf{w}^\star, b^\star)$ such that $\mathbf{a}^{\pm} = \mathbf{a}^\star$.

B.3.4 Symmetries arising from constant-odd activations

Now let $\sigma(z) = c + o(z)$ be constant-odd in the sense of Definition B.6. Using the same notation as before, the combined contribution of an aligned/opposite group with parameters $(\mathbf{w}, b)$ is

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \sum_k \mathbf{a}_k \, c + \sum_k \mathbf{a}_k \, o(z_k) = \mathbf{a}^{\pm} c + \mathbf{a}^{\mp} o(z). \tag{44}$$

Analogously to the even-linear case, $\mathbf{a}^{\pm}$ and $\mathbf{a}^{\mp}$ now control the constant and odd components, respectively. In particular, choosing $\mathbf{a}^{\mp} = \mathbf{0}$ generates a purely constant contribution, while an aligned/opposite group satisfying $\mathbf{a}^{\mp} = \mathbf{a}^\star$ reproduces the contribution of a single reference neuron with parameters $(\mathbf{w}, b, \mathbf{a}^\star)$ up to a residual constant contribution. This motivates the following two symmetry groups, corresponding to the "odd (+ constant)" symmetries of Martinelli et al. [24].

Definition B.11 (Constant-neuron group)**.** Suppose $\sigma(z) = c + o(z)$ is constant-odd. A *constant-neuron group* is a difference-nondegenerate aligned/opposite group with parameters $(\mathbf{w}, b)$ such that $\mathbf{a}^{\mp} = \mathbf{0}$.

Definition B.12 (Constant-duplicate-neuron group)**.** Suppose $\sigma(z) = c + o(z)$ is constant-odd, and fix parameters $(\mathbf{w}^\star, b^\star, \mathbf{a}^\star)$ with $\mathbf{a}^\star \neq \mathbf{0}$. A *constant-duplicate-neuron group* is a difference-nondegenerate aligned/opposite group with parameters $(\mathbf{w}^\star, b^\star)$ such that $\mathbf{a}^{\mp} = \mathbf{a}^\star$.

Note that constant-neuron *groups* are distinct from the *individual* constant neurons introduced in Definition B.4: the former involve aligned/opposite groups with nonzero incoming weights, whereas the latter are individual neurons with vanishing incoming weights.

B.3.5 Choice of sign in aligned/opposite groups

As noted after Definition B.7, the decomposition of an aligned/opposite group into subgroups $\mathcal{N}^+$ and $\mathcal{N}^-$ is defined only up to a global sign flip of the reference parameters $(\mathbf{w}, b)$.

For nonduplicate symmetry groups, this ambiguity is intrinsic and irrelevant. Linear-neuron groups and constant-neuron groups are defined by the conditions $\mathbf{a}^{\pm} = \mathbf{0}$ and $\mathbf{a}^{\mp} = \mathbf{0}$, respectively, which are symmetric under exchanging the aligned and opposite subgroups. In these cases, there is no intrinsic meaning attached to the labels "aligned" and "opposite": either choice yields the same symmetry group.

The duplicate variants are conceptually different. They are intended to replicate the behavior of a single reference neuron with parameters $(\mathbf{w}^\star, b^\star, \mathbf{a}^\star)$, which induces a distinguished notion of alignment *relative to that neuron*. This semantic "orientation" is present for both linear-duplicate and constant-duplicate groups. In the linear-duplicate case, the defining condition $\mathbf{a}^{\pm} = \mathbf{a}^\star$ happens to be

compatible with either choice of aligned/opposite decomposition, so that the sign ambiguity remains at the level of the formal definition. In contrast, for constant-duplicate groups, the defining condition $\mathbf{a}^{\mp} = \mathbf{a}^{\star} \neq \mathbf{0}$ selects a unique admissible decomposition, since reversing the sign would violate the defining equation.

Thus, while the aligned/opposite decomposition is a priori sign-ambiguous, this ambiguity either plays no role (for nonduplicate groups), matters semantically but not algebraically (for linear-duplicate groups), or is resolved by the defining equations themselves (for constant-duplicate groups).

B.3.6 Function-preserving realizations

Even-linear activations. For even-linear $\sigma(z) = e(z) + mz$, the combined contribution of an aligned/opposite group with parameters $(\mathbf{w}, b)$ equals $\mathbf{a}^{\pm} e(z) + \mathbf{a}^{\mp} mz$, where $z \coloneqq \mathbf{w}^{\top}\mathbf{x} + b$. A linear-neuron group ($\mathbf{a}^{\pm} = \mathbf{0}$, Definition B.9) thus contributes a linear term:

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \mathbf{a}^{\pm} e(z) + \mathbf{a}^{\mp} mz = \mathbf{a}^{\mp} mz. \tag{45}$$

Replacing an individual neuron with parameters $(\mathbf{w}^{\star}, b^{\star}, \mathbf{a}^{\star})$ by a linear-duplicate-neuron group ($\mathbf{a}^{\pm} = \mathbf{a}^{\star}$, Definition B.10) reproduces the reference neuron's contribution up to a residual linear term:

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \mathbf{a}^{\pm} e(z^{\star}) + \mathbf{a}^{\mp} mz^{\star} = \mathbf{a}^{\star} \, \sigma(z^{\star}) + (\mathbf{a}^{\mp} - \mathbf{a}^{\star}) \, mz^{\star}, \tag{46}$$

where $z^{\star} \coloneqq \mathbf{w}^{\star\top}\mathbf{x} + b^{\star}$. For the realized function to remain unchanged, these (residual) contributions must cancel collectively across linear-neuron and linear-duplicate-neuron groups and individual constant neurons.

Even activations. When σ is even, corresponding to the even-linear case with $m = 0$, both the linear contribution $\mathbf{a}^{\mp} mz$ of a linear-neuron group and the linear residual $(\mathbf{a}^{\mp} - \mathbf{a}^{\star}) \, mz^{\star}$ of a linear-duplicate-neuron group vanish. Consequently, linear-neuron groups and linear-duplicate-neuron groups are always function-preserving.

Constant-odd activations. For constant-odd $\sigma(z) = c + o(z)$, the combined contribution of an aligned/opposite group with parameters $(\mathbf{w}, b)$ is $\mathbf{a}^{\pm} c + \mathbf{a}^{\mp} o(z)$, with z as above. Adding a constant-neuron group ($\mathbf{a}^{\mp} = \mathbf{0}$, Definition B.11) thus contributes an input-independent offset:

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \mathbf{a}^{\pm} c + \mathbf{a}^{\mp} o(z) = \mathbf{a}^{\pm} c. \tag{47}$$

Replacing an individual neuron with parameters $(\mathbf{w}^{\star}, b^{\star}, \mathbf{a}^{\star})$ by a constant-duplicate-neuron group ($\mathbf{a}^{\mp} = \mathbf{a}^{\star}$, Definition B.12) reproduces the reference neuron's contribution up to a residual constant term:

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \mathbf{a}^{\pm} c + \mathbf{a}^{\mp} o(z^{\star}) = \mathbf{a}^{\star} \, \sigma(z^{\star}) + (\mathbf{a}^{\pm} - \mathbf{a}^{\star}) \, c, \tag{48}$$

with $z^{\star}$ as above. For the realized function to remain unchanged, these constant contributions must cancel collectively across constant-neuron and constant-duplicate-neuron groups and individual constant neurons (Definition B.4).

Odd activations. When σ is odd, corresponding to the constant-odd case with $c = 0$, both the constant contribution $\mathbf{a}^{\pm} c$ of a constant-neuron group and the constant residual $(\mathbf{a}^{\pm} - \mathbf{a}^{\star}) \, c$ of a constant-duplicate-neuron group vanish. Consequently, constant-neuron groups and constant-duplicate-neuron groups are always function-preserving.

Throughout, a finite collection of additions, replacements, removals, or collapses of the applicable activation-dependent groups and individual constant neurons is regarded as a single function-preserving transformation whenever its net residual contribution vanishes identically. The constituent operations need not preserve the function in isolation.

B.3.7 Minimality and nondegeneracy

Under our standing exclusion of affine activations, the structural conditions in Definition B.7, the class-specific nondegeneracy conditions in Definition B.8, and the requirement $\mathbf{a}^\star \neq \mathbf{0}$ in the duplicate variants jointly ensure that the four activation-dependent symmetry classes (Definitions B.9 to B.12) form a minimal taxonomy of mutually distinct classes disjoint from the activation-independent classes of Appendix B.2. Building on the definitions of Martinelli et al. [24], these conditions prevent activation-dependent groups from decomposing into smaller activation-dependent or activation-independent groups or degenerating into another activation-dependent class. This subsection treats each condition in turn, spelling out the degeneracy it rules out.

Nonzero shared incoming weights ($\mathbf{w} \neq \mathbf{0}$). If $\mathbf{w} = \mathbf{0}$, every neuron in the aligned/opposite group has incoming weights $\pm\mathbf{w} = \mathbf{0}$ and therefore qualifies as an individual constant neuron in the sense of Definition B.4. The combined contribution reduces to $\mathbf{a}^{\pm} e(b) + \mathbf{a}^{\mp} mb$ for an even-linear activation, and to $\mathbf{a}^{\pm} c + \mathbf{a}^{\mp} o(b)$ for a constant-odd activation, in either case indistinguishable from the contribution of an unstructured collection of constant neurons. Requiring $\mathbf{w} \neq \mathbf{0}$ thus ensures that activation-dependent symmetry groups capture genuinely input-dependent structure.

Two nonempty subgroups ($\mathcal{N}^+, \mathcal{N}^- \neq \emptyset$). If one of the subgroups were empty, all neurons would share the same incoming parameters. For the nonduplicate classes, $\mathbf{a}^{\pm} = \mathbf{0}$ (linear-neuron group) and $\mathbf{a}^{\mp} = \mathbf{0}$ (constant-neuron group) then both become $\sum_k \mathbf{a}_k = \mathbf{0}$, the defining condition of a zero-neuron group, whichever subgroup is empty. For the duplicate classes, by contrast, the resulting activation-independent symmetry depends on which subgroup is empty. If $\mathcal{N}^- = \emptyset$, both $\mathbf{a}^{\pm} = \mathbf{a}^\star$ (linear-duplicate-neuron group) and $\mathbf{a}^{\mp} = \mathbf{a}^\star$ (constant-duplicate-neuron group) become $\sum_k \mathbf{a}_k = \mathbf{a}^\star$, the defining condition of a duplicate-neuron group of the reference neuron. If $\mathcal{N}^+ = \emptyset$, they become $\sum_k \mathbf{a}_k = \mathbf{a}^\star$ and $\sum_k \mathbf{a}_k = -\mathbf{a}^\star$, respectively. The group then duplicates the flipped neuron $(-\mathbf{w}^\star, -b^\star, \mathbf{a}^\star)$ or $(-\mathbf{w}^\star, -b^\star, -\mathbf{a}^\star)$, whose contribution differs from that of the reference neuron by $-2m\,\mathbf{a}^\star z^\star$ or $-2c\,\mathbf{a}^\star$, respectively. Requiring both subgroups to be nonempty thus prevents activation-dependent symmetry groups from reducing to activation-independent ones, possibly applied to a flipped copy of the reference neuron.

Sum- and difference-nondegeneracy. For linear-neuron and linear-duplicate-neuron groups, the relevant aggregate of a subset $\mathcal{J} \subseteq \mathcal{K}$ is $\mathbf{a}^{\pm}_{\mathcal{J}}$, whereas for constant-neuron and constant-duplicate-neuron groups it is $\mathbf{a}^{\mp}_{\mathcal{J}}$. The applicable nondegeneracy condition requires this aggregate to be nonzero for every nonempty proper subset $\mathcal{J} \subsetneq \mathcal{K}$.

Suppose that this condition is violated and, among all nonempty proper subsets with vanishing relevant aggregate, choose the smallest such subset $\mathcal{J}$. If $\mathcal{J}$ lies entirely within $\mathcal{N}^+$ or $\mathcal{N}^-$, its relevant aggregate (i.e., $\mathbf{a}^{\pm}_{\mathcal{J}}$ or $\mathbf{a}^{\mp}_{\mathcal{J}}$) agrees, up to sign, with the sum of its readout weights. By the choice of $\mathcal{J}$, no nonempty proper subset of these readout weights sums to zero, so $\mathcal{J}$ is subset-nonzero (Definition B.1). Because their sum vanishes, the neurons indexed by $\mathcal{J}$ form a zero-neuron group. If instead $\mathcal{J}$ intersects both $\mathcal{N}^+$ and $\mathcal{N}^-$, the same choice ensures that it is sum-nondegenerate or difference-nondegenerate, as applicable. The neurons indexed by $\mathcal{J}$ therefore form a smaller linear-neuron or constant-neuron group. In either case, the complementary subset $\mathcal{K} \setminus \mathcal{J}$ retains the defining aggregate of the original group (i.e., $\mathbf{0}$ for a nonduplicate group and $\mathbf{a}^\star$ for a duplicate group) so that the original group is decomposable and hence not minimal.

Because $\mathcal{N}^+$ and $\mathcal{N}^-$ are themselves nonempty proper subsets of $\mathcal{K}$, either nondegeneracy condition also requires both these subgroups to have nonzero aggregate readout. Indeed,

$$\mathbf{a}^{\pm} = \mathbf{a}^{\mp} \iff \sum_{k \in \mathcal{N}^-} \mathbf{a}_k = \mathbf{0}, \qquad \mathbf{a}^{\pm} = -\mathbf{a}^{\mp} \iff \sum_{k \in \mathcal{N}^+} \mathbf{a}_k = \mathbf{0}. \tag{49}$$

If either equality held, the neurons in one entire subgroup could be decomposed into zero-neuron groups. In the nonduplicate case, the defining equation would force the aggregate readout of the other subgroup to vanish as well, so the full group would decompose into zero-neuron groups. In the duplicate case, the remaining subgroup consists entirely of neurons with identical incoming parameters and can therefore be reduced to a single neuron, possibly together with zero-neuron groups. Thus, in either case, the activation-dependent group would decompose into the activation-independent symmetry groups of Appendix B.2, possibly together with a flipped copy of the reference neuron.

Finally, in either duplicate class (i.e., linear or constant), a nonempty proper subset has the same relevant aggregate $\mathbf{a}^\star$ as the full group if and only if its nonempty complement has vanishing relevant aggregate. The applicable nondegeneracy condition therefore also excludes proper duplicate subgroups with respect to the same reference neuron, directly paralleling the minimality argument for activation-independent duplicate-neuron groups in Appendix B.2.2.

Nonzero reference readout ($\mathbf{a}^\star \neq \mathbf{0}$, duplicate variants). If $\mathbf{a}^\star = \mathbf{0}$, the defining conditions of the linear- and constant-duplicate-neuron groups, $\mathbf{a}^\pm = \mathbf{a}^\star$ and $\mathbf{a}^\mp = \mathbf{a}^\star$, would reduce to those of the linear- and constant-neuron groups, $\mathbf{a}^\pm = \mathbf{0}$ and $\mathbf{a}^\mp = \mathbf{0}$, making the duplicate variants indistinguishable from their nonduplicate counterparts. Moreover, a reference neuron with $\mathbf{a}^\star = \mathbf{0}$ contributes nothing to the realized function, so "duplicating" it would carry no semantic content. Requiring $\mathbf{a}^\star \neq \mathbf{0}$ keeps the duplicate and nonduplicate variants disjoint and ensures that duplicate variants genuinely replicate a nontrivial reference neuron.

B.3.8 Nonuniqueness of irreducible parameterizations

We now demonstrate that distinct irreducible parameterizations within the same symmetry orbit need not be related by generic reparameterization symmetries (Appendix B.1) alone.

Example B.13 (Oppositely oriented irreducible ReLU parameterizations)**.** Consider a one-hidden-layer network with one-dimensional input, one-dimensional output, and the ReLU activation

$$\psi(z) \coloneqq \max(0, z). \tag{50}$$

Let $\boldsymbol{\theta}^+$ be the width-3 parameterization consisting of the neurons

$$(w_1, b_1, a_1) = (1, 1, 1), \qquad (w_2, b_2, a_2) = (1, 0, -2), \qquad (w_3, b_3, a_3) = (1, -1, 1), \tag{51}$$

visualized in the left panel of Figure B.1. The three neurons have kinks at -1, 0, and 1, respectively, and together realize the *tent function*

$$f_{\boldsymbol{\theta}^+}(x) = \psi(x+1) - 2\,\psi(x) + \psi(x-1) = \psi(1 - |x|). \tag{52}$$

The same function is realized by the oppositely oriented parameterization $\boldsymbol{\theta}^-$ consisting of

$$\begin{aligned} (-w_1, -b_1, a_1) &= (-1, -1, 1), \\ (-w_2, -b_2, a_2) &= (-1, 0, -2), \\ (-w_3, -b_3, a_3) &= (-1, 1, 1), \end{aligned} \tag{53}$$

shown in the center panel of Figure B.1. Indeed, using $\psi(z) - \psi(-z) = z$, we obtain

$$\begin{aligned} &f_{\boldsymbol{\theta}^+}(x) - f_{\boldsymbol{\theta}^-}(x) \\ &= \big[\psi(x+1) - \psi(-x-1)\big] - 2\big[\psi(x) - \psi(-x)\big] + \big[\psi(x-1) - \psi(-x+1)\big] \\ &= (x+1) - 2x + (x-1) = 0. \end{aligned} \tag{54}$$

Moreover, these two parameterizations are symmetry-equivalent as defined in Section 3.2. To see this, write

$$z_1 \coloneqq x + 1, \qquad z_2 \coloneqq x, \qquad z_3 \coloneqq x - 1. \tag{55}$$

Define a width-6 parameterization $\boldsymbol{\eta}$ consisting, for each $j \in \{1, 2, 3\}$, of the two neurons

$$(w_j, b_j, -a_j) \qquad \text{and} \qquad (-w_j, -b_j, a_j). \tag{56}$$

Since ReLU is even-linear (see Appendix C.1) and $a_j \neq 0$, each such pair is sum-nondegenerate. Its aggregate readout vanishes, $a^\pm = -a_j + a_j = 0$, so each pair forms a linear-neuron group. The contribution of the jth group to the function $f_{\boldsymbol{\eta}}$ realized by $\boldsymbol{\eta}$ is

$$-a_j\,\psi(z_j) + a_j\,\psi(-z_j) = -a_j z_j. \tag{57}$$

Collectively, these contributions cancel:

$$f_{\boldsymbol{\eta}}(x) = -\sum_{j=1}^{3} a_j z_j = -(x+1) + 2x - (x-1) = 0. \tag{58}$$

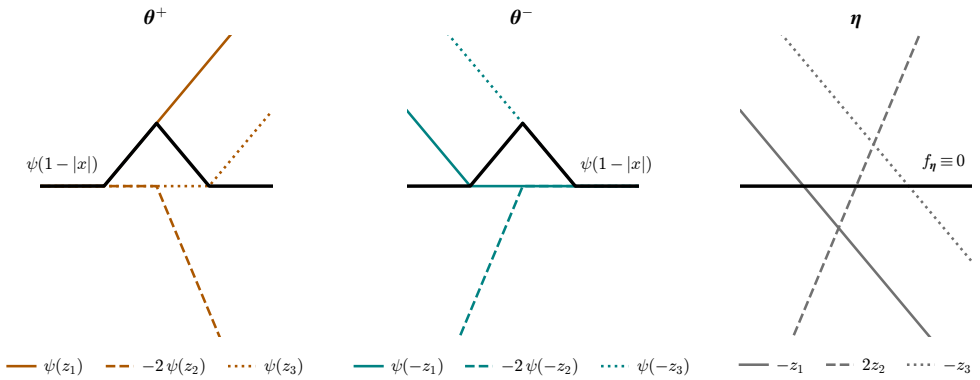

Figure B.1: **Oppositely oriented irreducible parameterizations of the tent function**. **(Left)** A parameterization $\boldsymbol{\theta}^+$ of the tent function $\psi(1-|x|) = \max(0, 1-|x|)$ composed of three ReLU neurons (w_j, b_j, a_j), all with positive incoming weight $w_j > 0$. **(Center)** The same function, realized by an implementation $\boldsymbol{\theta}^-$ consisting of the three neurons $(-w_j, -b_j, a_j)$ with negated incoming parameters. **(Right)** The net contributions $-a_j z_j$ of the three linear-neuron groups in the parameterization $\boldsymbol{\eta}$ used to demonstrate symmetry-equivalence between $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$. No linear-neuron group vanishes individually, but their contributions sum to zero, so that $f_{\boldsymbol{\eta}}$ realizes the zero function.

Thus, $\boldsymbol{\eta}$ consists entirely of symmetry groups and realizes the zero function. The parameterization $\boldsymbol{\theta}^+$ is therefore symmetry-equivalent to the parameterization $\boldsymbol{\theta}^+ \oplus \boldsymbol{\eta}$ obtained by combining all neurons from $\boldsymbol{\theta}^+$ and $\boldsymbol{\eta}$. In this combined parameterization, each original neuron (w_j, b_j, a_j) of $\boldsymbol{\theta}^+$ and its readout-negated copy $(w_j, b_j, -a_j)$ of $\boldsymbol{\eta}$ form a zero-neuron group. Removing these three groups leaves only the oppositely oriented neurons of $\boldsymbol{\eta}$, which are exactly the neurons of $\boldsymbol{\theta}^-$, showing that $\boldsymbol{\theta}^+ \sim \boldsymbol{\theta}^-$.

The realized tent function $\psi(1-|x|)$ has three distinct kinks with nonzero changes in slope, whereas each one-dimensional ReLU neuron can introduce at most one such kink. Any ReLU network realizing the tent function must therefore have width at least 3. Since both $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ attain this minimum, it follows that both parameterizations are irreducible.

Finally, the two parameterizations cannot be related by neuron permutations and positive scaling. Every neuron in $\boldsymbol{\theta}^+$ has positive incoming weight, whereas every neuron in $\boldsymbol{\theta}^-$ has negative incoming weight, and neither permutation nor positive scaling can reverse this orientation.

Example B.13 illustrates three important facts about symmetry-equivalent parameterizations.

- The activation-dependent symmetry groups discussed in this subsection need not be function-preserving individually. In fact, none of the linear-neuron groups in Example B.13 is function-preserving when added on its own; only their joint addition yields a function-preserving transformation.
- Irreducible parameterizations are *not* unique up to generic reparameterization symmetries (Appendix B.1) when the activation admits any of the activation-dependent symmetries discussed in this subsection.
- Finally, the construction of the parameterization $\boldsymbol{\eta}$ used to establish symmetry-equivalence between the oppositely oriented parameterizations $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ generalizes to yield a quantitative criterion characterizing when two parameterizations are symmetry-equivalent (see the proof of Proposition B.15).

B.4 Parameter classes, orbit invariants, and symmetry-equivalence

The definition of symmetry-equivalence established in Section 3.2 of the main text is not directly operational: given parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, we lack a criterion for deciding whether the two are symmetry-equivalent. To establish such a criterion, we first introduce *parameter classes* and rewrite

the realized function $f_{\boldsymbol{\theta}}$ in terms of them (Appendix B.4.1). We then show that these classes yield a quantitative, easily checkable criterion for symmetry-equivalence (Appendix B.4.2).

Terminology. Positively homogeneous activations of degree 1 form an important subclass of even-linear activations (Corollary C.2). Because their inherent positive-scaling symmetry affects the theory developed in this subsection, they must be treated separately from even-linear activations that are not positively 1-homogeneous. Accordingly, throughout this subsection, we reserve the term *even-linear* exclusively for even-linear activations that are *not* positively 1-homogeneous.

B.4.1 Parameter classes

As in the main text, we write

$$\overline{\mathbf{w}} \coloneqq (\mathbf{w}^\top, b)^\top, \qquad \overline{\mathbf{x}} \coloneqq (\mathbf{x}^\top, 1)^\top, \qquad z \coloneqq \overline{\mathbf{w}}^\top \overline{\mathbf{x}}. \tag{59}$$

Definition B.14 (Parameter classes)**.** Fix a nonaffine activation σ and an incoming-parameter vector $\overline{\mathbf{w}} = (\mathbf{w}^\top, b)^\top$ with $\mathbf{w} \neq \mathbf{0}$. Its *parameter class* is

$$[\overline{\mathbf{w}}] \coloneqq \begin{cases} \{\alpha\overline{\mathbf{w}} \mid \alpha \in \mathbb{R}_{\neq 0}\}, & \sigma \text{ is positively homogeneous of degree 1,} \\ \{\overline{\mathbf{w}}, -\overline{\mathbf{w}}\}, & \sigma \text{ is even-linear or constant-odd,} \\ \{\overline{\mathbf{w}}\}, & \text{otherwise.} \end{cases} \tag{60}$$

Mirroring Section 3.3, we denote the parameter classes represented by a parameterization $\boldsymbol{\theta}$ by

$$\mathcal{Q}(\boldsymbol{\theta}) \coloneqq \{[\overline{\mathbf{w}}_j] \mid \mathbf{w}_j \neq \mathbf{0}\}, \tag{61}$$

and we let $\mathcal{J}_q \coloneqq \{j \mid \overline{\mathbf{w}}_j \in q\}$ index the neurons of class $q \in \mathcal{Q}(\boldsymbol{\theta})$.

For every $q \in \mathcal{Q}(\boldsymbol{\theta})$, we fix a representative $\overline{\mathbf{w}}_q \in q$. When σ is positively homogeneous of degree 1, we choose $\overline{\mathbf{w}}_q$ to have unit norm, $\|\overline{\mathbf{w}}_q\| = 1$. When σ is constant-odd, we choose the unique element of q whose incoming-weight vector has a positive first nonzero entry. We write

$$\mathcal{J}_q^+ \coloneqq \{j \in \mathcal{J}_q \mid \overline{\mathbf{w}}_j = \alpha\overline{\mathbf{w}}_q \text{ for some } \alpha > 0\}, \tag{62}$$

and, analogously,

$$\mathcal{J}_q^- \coloneqq \{j \in \mathcal{J}_q \mid \overline{\mathbf{w}}_j = -\alpha\overline{\mathbf{w}}_q \text{ for some } \alpha > 0\}, \tag{63}$$

so that $\mathcal{J}_q = \mathcal{J}_q^+ \sqcup \mathcal{J}_q^-$. Finally, as in the main text, let $\mathcal{J}_0 \coloneqq \{j \mid \mathbf{w}_j = \mathbf{0}\}$ index all constant neurons of a parameterization.

Positively homogeneous activations of degree 1. Let σ be a positively homogeneous activation of degree 1. By Corollary C.2, we can write $\sigma(z) = \delta|z| + mz$, where

$$\delta \coloneqq \frac{\lambda_+ - \lambda_-}{2}, \qquad m \coloneqq \frac{\lambda_+ + \lambda_-}{2}, \qquad \lambda_-, \lambda_+ \in \mathbb{R}. \tag{64}$$

For $q = \{\alpha\overline{\mathbf{w}} \mid \alpha \in \mathbb{R}_{\neq 0}\} \in \mathcal{Q}(\boldsymbol{\theta})$, define

$$\phi_q(\mathbf{x}) \coloneqq \delta\frac{|z|}{\|\overline{\mathbf{w}}\|}, \qquad \boldsymbol{\beta}_q(\boldsymbol{\theta}) \coloneqq \sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\|\, \mathbf{a}_j, \tag{65}$$

and

$$\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq m \sum_{j \notin \mathcal{J}_0} \mathbf{a}_j z_j + \sum_{j \in \mathcal{J}_0} \mathbf{a}_j \sigma(b_j). \tag{66}$$

Note that $\phi_q \colon \mathbb{R}^{N_i} \to \mathbb{R}$ depends only on the parameter class $q = \{\alpha\overline{\mathbf{w}} \mid \alpha \in \mathbb{R}_{\neq 0}\}$, *not* on the particular representative, since $|z|/\|\overline{\mathbf{w}}\|$ is unchanged when $\overline{\mathbf{w}}$ is replaced by $\alpha\overline{\mathbf{w}}$ for any $\alpha \neq 0$.

Even-linear activations. Let $\sigma(z) = e(z) + mz$, with e even. For $q = \{\overline{\mathbf{w}}, -\overline{\mathbf{w}}\} \in \mathcal{Q}(\boldsymbol{\theta})$, define

$$\phi_q(\mathbf{x}) \coloneqq e(\overline{\mathbf{w}}^\top \overline{\mathbf{x}}), \qquad \boldsymbol{\beta}_q(\boldsymbol{\theta}) \coloneqq \sum_{j \in \mathcal{J}_q} \mathbf{a}_j, \tag{67}$$

and

$$\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq m \sum_{j \notin \mathcal{J}_0} \mathbf{a}_j z_j + \sum_{j \in \mathcal{J}_0} \mathbf{a}_j \sigma(b_j). \tag{68}$$

Again, the definition of ϕ_q is independent of the choice of representative, since $e(-\overline{\mathbf{w}}^\top \overline{\mathbf{x}}) = e(\overline{\mathbf{w}}^\top \overline{\mathbf{x}})$.

Constant-odd activations. Let $\sigma(z) = c + o(z)$, with o odd. For $q = \{\overline{\mathbf{w}}_q, -\overline{\mathbf{w}}_q\} \in \mathcal{Q}(\boldsymbol{\theta})$, define

$$\phi_q(\mathbf{x}) \coloneqq o(\overline{\mathbf{w}}_q^\top \overline{\mathbf{x}}), \qquad \boldsymbol{\beta}_q(\boldsymbol{\theta}) \coloneqq \sum_{j \in \mathcal{J}_q^+} \mathbf{a}_j - \sum_{j \in \mathcal{J}_q^-} \mathbf{a}_j, \tag{69}$$

and

$$\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq c \sum_{j \notin \mathcal{J}_0} \mathbf{a}_j + \sum_{j \in \mathcal{J}_0} \mathbf{a}_j \sigma(b_j). \tag{70}$$

For constant-odd activations, unlike the positively 1-homogeneous and even-linear cases, the definitions of $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ and ϕ_q depend on the chosen representative of q. Replacing $\overline{\mathbf{w}}_q$ by $-\overline{\mathbf{w}}_q$ exchanges $\mathcal{J}_q^+$ and $\mathcal{J}_q^-$ in Equations (62) and (63), thereby negating $\boldsymbol{\beta}_q(\boldsymbol{\theta})$. Under the same replacement, ϕ_q changes sign because $o(-\overline{\mathbf{w}}_q^\top \overline{\mathbf{x}}) = -o(\overline{\mathbf{w}}_q^\top \overline{\mathbf{x}})$. Hence, the product $\boldsymbol{\beta}_q(\boldsymbol{\theta})\phi_q(\mathbf{x})$ is independent of the representative, but fixing $\overline{\mathbf{w}}_q$ is necessary to define the two factors individually.

Activations with neither symmetry property. Let σ be neither even-linear nor constant-odd. For $q = \{\overline{\mathbf{w}}\} \in \mathcal{Q}(\boldsymbol{\theta})$, define

$$\phi_q(\mathbf{x}) \coloneqq \sigma(\overline{\mathbf{w}}^\top \overline{\mathbf{x}}), \qquad \boldsymbol{\beta}_q(\boldsymbol{\theta}) \coloneqq \sum_{j \in \mathcal{J}_q} \mathbf{a}_j, \tag{71}$$

and

$$\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq \sum_{j \in \mathcal{J}_0} \mathbf{a}_j \sigma(b_j), \tag{72}$$

which is constant in $\mathbf{x} \in \mathbb{R}^{N_i}$.

In all four cases, the definitions of $\boldsymbol{\beta}_q(\boldsymbol{\theta})$, $\phi_q(\mathbf{x})$, and $\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x})$ in Equations (65) to (72) allow us to express the function $f_{\boldsymbol{\theta}}$ realized by the parameterization $\boldsymbol{\theta}$ as follows:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{q \in \mathcal{Q}(\boldsymbol{\theta})} \boldsymbol{\beta}_q(\boldsymbol{\theta})\, \phi_q(\mathbf{x}) + \mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^{N_i}. \tag{73}$$

B.4.2 Characterization of symmetry-equivalence

We now extend the characterization of symmetry-equivalent parameterizations from the main text to all activation classes considered here.

Proposition B.15 (Characterization of symmetry-equivalence)**.** *Two parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, possibly of different widths, with the same nonaffine activation σ, are symmetry-equivalent if and only if*

$$\boldsymbol{\beta}_q(\boldsymbol{\theta}) = \boldsymbol{\beta}_q(\boldsymbol{\xi}) \quad \textit{for every } q \in \mathcal{Q}(\boldsymbol{\theta}) \cup \mathcal{Q}(\boldsymbol{\xi}), \qquad \mathbf{r}_{\boldsymbol{\theta}} = \mathbf{r}_{\boldsymbol{\xi}}, \tag{74}$$

with $\boldsymbol{\beta}_q(\cdot)$ understood to be zero for an absent class.

Proof. Suppose first that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are symmetry-equivalent, $\boldsymbol{\theta} \sim \boldsymbol{\xi}$.

Equality of the residuals $\mathbf{r}_{\boldsymbol{\theta}}$ and $\mathbf{r}_{\boldsymbol{\xi}}$. We first show that equality of the aggregates $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ and $\boldsymbol{\beta}_q(\boldsymbol{\xi})$ implies equality of the residual functions $\mathbf{r}_{\boldsymbol{\theta}}$ and $\mathbf{r}_{\boldsymbol{\xi}}$. Assume that $\boldsymbol{\beta}_q(\boldsymbol{\theta}) = \boldsymbol{\beta}_q(\boldsymbol{\xi})$ for every parameter class $q \in \mathcal{Q}(\boldsymbol{\theta}) \cup \mathcal{Q}(\boldsymbol{\xi})$. Since symmetry-equivalence implies $f_{\boldsymbol{\theta}} = f_{\boldsymbol{\xi}}$, the decomposition in Equation (73) gives

$$\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{r}_{\boldsymbol{\xi}}(\mathbf{x}) = \underbrace{f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\xi}}(\mathbf{x})}_{=\mathbf{0}} - \sum_q \underbrace{(\boldsymbol{\beta}_q(\boldsymbol{\theta}) - \boldsymbol{\beta}_q(\boldsymbol{\xi}))}_{=\mathbf{0}} \phi_q(\mathbf{x}) = \mathbf{0}, \qquad \mathbf{x} \in \mathbb{R}^{N_i}, \tag{75}$$

and hence $\mathbf{r}_{\boldsymbol{\theta}} = \mathbf{r}_{\boldsymbol{\xi}}$.

Equality of the class aggregates $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ and $\boldsymbol{\beta}_q(\boldsymbol{\xi})$. It remains to show that symmetry-equivalence implies equality of $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ and $\boldsymbol{\beta}_q(\boldsymbol{\xi})$ for all parameter classes q. Since the parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are related through a finite composition of the symmetries in Appendices B.1 to B.3, it suffices to show that $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ is invariant under each such symmetry.

The permutation symmetry (Appendix B.1.1) simply permutes neurons and has no effect on the aggregates $\boldsymbol{\beta}_q(\boldsymbol{\theta})$. The positive-scaling symmetry (Appendix B.1.2) of positively 1-homogeneous activations preserves both the parameter class q and its contribution to the corresponding class aggregate $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ in Equation (65), since

$$\|\alpha\overline{\mathbf{w}}_j\|\,\alpha^{-1}\mathbf{a}_j = \|\overline{\mathbf{w}}_j\|\,\mathbf{a}_j, \qquad \alpha > 0. \tag{76}$$

For odd activations, the sign-flip symmetry $(\overline{\mathbf{w}}_j, \mathbf{a}_j) \mapsto (-\overline{\mathbf{w}}_j, -\mathbf{a}_j)$ (Appendix B.1.3) moves the jth neuron from $\mathcal{J}_q^+$ to $\mathcal{J}_q^-$, or vice versa, while negating its readout $\mathbf{a}_j$. These two changes cancel in the signed difference defining $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ in Equation (69), leaving the aggregate invariant. For even activations, the sign-flip symmetry $(\overline{\mathbf{w}}_j, \mathbf{a}_j) \mapsto (-\overline{\mathbf{w}}_j, \mathbf{a}_j)$ (Appendix B.1.3) keeps neuron j in the same parameter class q while leaving its readout $\mathbf{a}_j$ unchanged. Its contribution to the aggregate $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ in Equation (67) is therefore unchanged as well.

In a zero-neuron group (Definition B.2), all neurons share the same incoming-parameter vector $\overline{\mathbf{w}}$ and therefore belong to the same parameter class q. Because their aggregate readout vanishes, $\sum_z \mathbf{a}_z = \mathbf{0}$, adding or removing a zero-neuron group does not change $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ or any other class aggregate. Similarly, replacing a neuron by a duplicate-neuron group (Definition B.3), or collapsing such a group back to a single neuron, preserves both the parameter class q and its aggregate $\boldsymbol{\beta}_q(\boldsymbol{\theta})$. Individual constant neurons (Definition B.4), by contrast, contribute only to the residual $\mathbf{r}_{\boldsymbol{\theta}}$ and therefore do not affect any aggregate $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ to begin with.

Finally, an aligned/opposite group (Definition B.7) with incoming-parameter vector $\overline{\mathbf{w}}$ affects only the class aggregate $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ associated with parameter class $q = [\overline{\mathbf{w}}]$. For positively homogeneous activations of degree 1, the group's contribution to this aggregate is

$$\sum_{k\in\mathcal{K}} \|\overline{\mathbf{w}}_k\|\,\mathbf{a}_k = \|\overline{\mathbf{w}}\| \sum_{k\in\mathcal{K}} \mathbf{a}_k = \|\overline{\mathbf{w}}\|\,\mathbf{a}^{\pm}, \tag{77}$$

where $\mathcal{K}$ indexes the neurons of the aligned/opposite group. For even-linear activations, the corresponding contribution is

$$\sum_{k\in\mathcal{K}} \mathbf{a}_k = \mathbf{a}^{\pm}, \tag{78}$$

and for constant-odd activations, it is

$$\varepsilon \sum_{k\in\mathcal{N}^+} \mathbf{a}_k - \varepsilon \sum_{k\in\mathcal{N}^-} \mathbf{a}_k = \varepsilon\mathbf{a}^{\mp}, \tag{79}$$

where $\varepsilon \in \{-1, 1\}$ is such that $\varepsilon\overline{\mathbf{w}} = \overline{\mathbf{w}}_q$. Linear-neuron groups (Definition B.9) and constant-neuron groups (Definition B.11) satisfy $\mathbf{a}^{\pm} = \mathbf{0}$ and $\mathbf{a}^{\mp} = \mathbf{0}$, respectively, so adding or removing such groups leaves $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ unchanged. Likewise, linear-duplicate-neuron groups (Definition B.10) and constant-duplicate-neuron groups (Definition B.12) satisfy $\mathbf{a}^{\pm} = \mathbf{a}^{\star}$ and $\mathbf{a}^{\mp} = \mathbf{a}^{\star}$, respectively, where $\mathbf{a}^{\star}$ denotes the reference neuron's readout. Thus, replacing a neuron by such a duplicate group, or collapsing the group back to a single neuron, again leaves $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ invariant.

Conversely, suppose that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ satisfy $\boldsymbol{\beta}_q(\boldsymbol{\theta}) = \boldsymbol{\beta}_q(\boldsymbol{\xi})$ for all parameter classes $q \in \mathcal{Q}(\boldsymbol{\theta}) \cup \mathcal{Q}(\boldsymbol{\xi})$, and that $\mathbf{r}_{\boldsymbol{\theta}} = \mathbf{r}_{\boldsymbol{\xi}}$. To prove that such parameterizations are symmetry-equivalent, we generalize the construction of the parameterization $\boldsymbol{\eta}$ from Example B.13, used there to establish equivalence between two oppositely oriented ReLU parameterizations $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ realizing the tent function. Specifically, we construct a parameterization $\boldsymbol{\eta}$ realizing the zero function, $f_{\boldsymbol{\eta}} \equiv \mathbf{0}$, whose neurons can be partitioned into zero-neuron groups, linear-neuron groups or constant-neuron groups (depending on the activation σ), and individual constant neurons. The parameterization obtained by combining the neurons of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ then places every neuron of $\boldsymbol{\theta}$ into a removable symmetry group, leaving precisely the neurons of $\boldsymbol{\xi}$, as we shall see.

The auxiliary parameterization $\boldsymbol{\eta}$. First, let $\boldsymbol{\eta}$ denote the parameterization formed from all neurons in $\boldsymbol{\theta}$ with their readouts negated, together with all neurons in $\boldsymbol{\xi}$. If σ is positively homogeneous of

degree 1, the positive-scaling symmetry allows each nonconstant neuron in $\boldsymbol{\eta}$ to be transformed as

$$(\overline{\mathbf{w}}_j, \mathbf{a}_j) \mapsto (\|\overline{\mathbf{w}}_j\|^{-1}\overline{\mathbf{w}}_j, \|\overline{\mathbf{w}}_j\|\,\mathbf{a}_j). \tag{80}$$

Because positive scaling preserves symmetry-equivalence, parameter classes, class aggregates, and residuals, we may assume without loss of generality that $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ have been normalized in this way before constructing $\boldsymbol{\eta}$ whenever σ is positively 1-homogeneous. The resulting auxiliary parameterization $\boldsymbol{\eta}$ is then normalized as well. In that case, within each parameter class $q = [\overline{\mathbf{w}}]$, the incoming-parameter vector of every neuron equals one of the two unit-norm orientations $\pm\|\overline{\mathbf{w}}\|^{-1}\overline{\mathbf{w}}$. The positively 1-homogeneous and even-linear cases can then be treated identically below.

By construction of $\boldsymbol{\eta}$,

$$\mathcal{Q}(\boldsymbol{\eta}) = \mathcal{Q}(\boldsymbol{\theta}) \cup \mathcal{Q}(\boldsymbol{\xi}). \tag{81}$$

Since both the class aggregates $\boldsymbol{\beta}_q$ and the residual $\mathbf{r}$ are linear in the readouts $\mathbf{a}_j$, we have

$$\boldsymbol{\beta}_q(\boldsymbol{\eta}) = \boldsymbol{\beta}_q(\boldsymbol{\xi}) - \boldsymbol{\beta}_q(\boldsymbol{\theta}) = \mathbf{0} \quad \text{for every } q \in \mathcal{Q}(\boldsymbol{\eta}), \qquad \mathbf{r}_{\boldsymbol{\eta}} = \mathbf{r}_{\boldsymbol{\xi}} - \mathbf{r}_{\boldsymbol{\theta}} = \mathbf{0}. \tag{82}$$

Hence, by the decomposition in Equation (73), $\boldsymbol{\eta}$ realizes the zero function:

$$f_{\boldsymbol{\eta}}(\mathbf{x}) = \sum_{q \in \mathcal{Q}(\boldsymbol{\eta})} \boldsymbol{\beta}_q(\boldsymbol{\eta})\,\phi_q(\mathbf{x}) + \mathbf{r}_{\boldsymbol{\eta}}(\mathbf{x}) = \mathbf{0}, \qquad \mathbf{x} \in \mathbb{R}^{N_i}. \tag{83}$$

Decomposing $\boldsymbol{\eta}$ into symmetry groups. Next, we show that all nonconstant neurons in $\boldsymbol{\eta}$ can be organized into symmetry groups. Fix a parameter class $q \in \mathcal{Q}(\boldsymbol{\eta})$ and choose a nonempty subset $\mathcal{K} \subseteq \mathcal{J}_q$ of minimum cardinality whose contribution to $\boldsymbol{\beta}_q(\boldsymbol{\eta})$ vanishes. Such a subset exists because $\boldsymbol{\beta}_q(\boldsymbol{\eta}) = \mathbf{0}$. Suppose first that $\mathcal{K}$ contains only neurons whose incoming-parameter vectors are identical. Then, their vanishing contribution to $\boldsymbol{\beta}_q(\boldsymbol{\eta})$ implies $\sum_{k\in\mathcal{K}} \mathbf{a}_k = \mathbf{0}$. By minimality of $\mathcal{K}$, the collection $\{\mathbf{a}_k\}_{k\in\mathcal{K}}$ is subset-nonzero (Definition B.1), so the neurons indexed by $\mathcal{K}$ form a zero-neuron group (Definition B.2). Now suppose that $\mathcal{K}$ contains neurons of both orientations, which can only arise when σ is even-linear (including the positively 1-homogeneous case after the normalization above) or constant-odd. In this case, the neurons in $\mathcal{K}$ form an aligned/opposite group (Definition B.7). For even-linear activations, the vanishing contribution of $\mathcal{K}$ to $\boldsymbol{\beta}_q(\boldsymbol{\eta})$ is equivalent to $\mathbf{a}_{\mathcal{K}}^{\pm} = \mathbf{0}$. Minimality of $\mathcal{K}$ implies sum-nondegeneracy (Definition B.8), so the neurons in $\mathcal{K}$ form a linear-neuron group (Definition B.9). For constant-odd activations, the corresponding condition is $\mathbf{a}_{\mathcal{K}}^{\mp} = \mathbf{0}$, and minimality implies difference-nondegeneracy (Definition B.8). Hence, the neurons in $\mathcal{K}$ form a constant-neuron group (Definition B.11). In every case, $\mathcal{K}$ indexes a symmetry group whose contribution to $\boldsymbol{\beta}_q(\boldsymbol{\eta})$ vanishes. Removing this group therefore leaves a smaller set $\mathcal{J}_q \setminus \mathcal{K}$ of neurons in q whose contribution to the class aggregate $\boldsymbol{\beta}_q(\boldsymbol{\eta})$ still vanishes. Since $\mathcal{J}_q$ is finite, repeating the argument partitions all neurons in q into symmetry groups. Consequently, applying the same construction to every $q \in \mathcal{Q}(\boldsymbol{\eta})$ partitions all nonconstant neurons in $\boldsymbol{\eta}$ into zero-neuron groups, together with linear-neuron groups for even-linear activations or constant-neuron groups for constant-odd activations.

For parameterizations $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ with the same activation, we write $\boldsymbol{\theta}_1 \oplus \boldsymbol{\theta}_2$ for the parameterization obtained by combining their neurons and, when $\boldsymbol{\theta}_2$ consists of a subset of the neurons of $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_1 \ominus \boldsymbol{\theta}_2$ for the parameterization obtained by removing those neurons from $\boldsymbol{\theta}_1$.

$\boldsymbol{\theta} \sim \boldsymbol{\theta} \oplus \boldsymbol{\eta}$. We now show that $\boldsymbol{\theta}$ is symmetry-equivalent to $\boldsymbol{\theta} \oplus \boldsymbol{\eta}$. Denote the parameterization consisting of all zero-neuron groups in $\boldsymbol{\eta}$ by $\boldsymbol{\eta}_0$. Since adding zero-neuron groups is function-preserving (Appendix B.2.1), $\boldsymbol{\theta} \sim \boldsymbol{\theta} \oplus \boldsymbol{\eta}_0$. Because every zero-neuron group has vanishing aggregate readout, $\boldsymbol{\eta} \ominus \boldsymbol{\eta}_0$ has the same class aggregates and residual as $\boldsymbol{\eta}$, and hence still realizes the zero function by Equation (83). By the preceding decomposition of $\boldsymbol{\eta}$, $\boldsymbol{\eta} \ominus \boldsymbol{\eta}_0$ consists only of linear-neuron groups or constant-neuron groups, when admitted by the activation σ, together with individual constant neurons. Thus,

$$\boldsymbol{\theta} \sim \boldsymbol{\theta} \oplus \boldsymbol{\eta}_0 \sim (\boldsymbol{\theta} \oplus \boldsymbol{\eta}_0) \oplus (\boldsymbol{\eta} \ominus \boldsymbol{\eta}_0) = \boldsymbol{\theta} \oplus \boldsymbol{\eta}. \tag{84}$$

$\boldsymbol{\theta} \oplus \boldsymbol{\eta} \sim \boldsymbol{\xi}$. It remains to show that $\boldsymbol{\theta} \oplus \boldsymbol{\eta} \sim \boldsymbol{\xi}$. By construction of $\boldsymbol{\eta}$, the parameterization $\boldsymbol{\theta} \oplus \boldsymbol{\eta}$ contains all neurons of $\boldsymbol{\xi}$ together with, for every neuron $(\overline{\mathbf{w}}_j, \mathbf{a}_j)$ of $\boldsymbol{\theta}$, both that neuron and a copy $(\overline{\mathbf{w}}_j, -\mathbf{a}_j)$ with negated readout. If $\mathbf{w}_j = \mathbf{0}$, the pair consists of two constant neurons whose contributions $\mathbf{a}_j\sigma(b_j)$ and $-\mathbf{a}_j\sigma(b_j)$ cancel. If $\mathbf{w}_j \neq \mathbf{0}$ and $\mathbf{a}_j = \mathbf{0}$, each neuron is a singleton

zero-neuron group. If $\mathbf{w}_j \neq \mathbf{0}$ and $\mathbf{a}_j \neq \mathbf{0}$, the two neurons together form a zero-neuron group. Thus, every such pair can be removed from $\boldsymbol{\theta} \oplus \boldsymbol{\eta}$ while maintaining symmetry-equivalence. This leaves precisely the neurons of $\boldsymbol{\xi}$, so $\boldsymbol{\theta} \oplus \boldsymbol{\eta} \sim \boldsymbol{\xi}$. □

Proposition B.15 establishes that the aggregates $\boldsymbol{\beta}_q(\boldsymbol{\theta})$ and the residual function $\mathbf{r}_{\boldsymbol{\theta}}$ are invariants of the symmetry orbit containing $\boldsymbol{\theta}$. We therefore extend the shorthand notation introduced in the main text for positively 1-homogeneous activations to all activation classes considered here, suppressing the dependence on the particular parameterization $\boldsymbol{\theta}$.

Definition B.16 (Essential parameter classes)**.** Fix a symmetry orbit $\mathcal{O}$ and let $\boldsymbol{\theta} \in \mathcal{O}$ be any parameterization. We write

$$\boldsymbol{\beta}_q \coloneqq \boldsymbol{\beta}_q(\boldsymbol{\theta}), \qquad q \in \mathcal{Q}(\boldsymbol{\theta}), \tag{85}$$

and

$$\mathbf{r} \coloneqq \mathbf{r}_{\boldsymbol{\theta}}, \qquad \mathcal{E} \coloneqq \{q \in \mathcal{Q}(\boldsymbol{\theta}) \mid \boldsymbol{\beta}_q \neq \mathbf{0}\}. \tag{86}$$

A parameter class q is called *essential* if $q \in \mathcal{E}$.

C Symmetry properties of common activation functions

Which of the overparameterization symmetries introduced in Appendix B can arise in a given network depends on algebraic properties of its activation function σ. Here, we classify every scalar activation available from the `jax.nn` module [v0.10.0, 54], covering the standard activations commonly used in practice, according to whether it is even-linear (Appendix C.1), constant-odd (Appendix C.2), positively homogeneous of degree 1 (Appendix C.3), or none of the above (Appendix C.4). These classes are not mutually exclusive, and the complete classification is summarized in Table C.1. For completeness, this classification includes the identity, although affine activations are excluded from the nonlinear-network setting we consider.

C.1 Even-linear activations

We verify that each of the following `jax.nn` activations is even-linear in the sense of Definition B.5, and we report the slope m of its linear component.

GELU. For $\sigma(z) = z\,\Phi(z)$,[4] where $\Phi(z) = \frac{1}{2}(1 + \operatorname{erf}(z/\sqrt{2}))$ is the standard normal cumulative distribution function, the symmetry $\Phi(-z) = 1 - \Phi(z)$ gives

$$\sigma(-z) = -z\,\Phi(-z) = -z(1 - \Phi(z)). \tag{87}$$

The even component is not constant,

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{z\,\Phi(z) - z(1 - \Phi(z))}{2} = z\,\Phi(z) - \frac{z}{2}, \tag{88}$$

so GELU is not constant-odd. On the other hand, the odd component is linear,

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{z\,\Phi(z) + z(1 - \Phi(z))}{2} = \frac{z}{2}, \tag{89}$$

implying that GELU is even-linear with $m = \frac{1}{2}$.

Hard SiLU. Since the hard sigmoid satisfies $\mathrm{hard_sigmoid}(z) = \frac{1}{2} + o(z)$ for some odd function o, and $\mathrm{hard_silu}(z) = z\ \mathrm{hard_sigmoid}(z)$, we have

$$\mathrm{hard_silu}(z) = z\,o(z) + \frac{z}{2}, \tag{90}$$

where the product $z\,o(z)$ is even and $\frac{z}{2}$ is odd. The product $z\,o(z)$ is not constant (e.g., $z\,o(z) = \frac{z}{2}$ for $z \geq 3$), so the hard SiLU is not constant-odd. The odd component $\frac{z}{2}$ is linear, and the hard SiLU activation function is even-linear with $m = \frac{1}{2}$.

Identity. The identity $\sigma(z) = z$ is linear, so it is trivially even-linear with slope $m = 1$.

[4] Here, GELU refers to the exact formulation rather than the approximation used by default in `jax.nn`.

Table C.1: Symmetries of all scalar activation functions in `jax.nn` (v0.10.0). The aliases `swish` and `hard_swish` for `silu` and `hard_silu`, respectively, are omitted. We assume $\alpha > 0$ (`celu`, `elu`, `leaky_relu`, `selu`), $\lambda \geq 1$ (`selu`), and $b > 0$ (`squareplus`). For `leaky_relu`, we additionally assume $\alpha \neq 1$, since $\alpha = 1$ gives the identity. **Even+Lin** and **Const+Odd** report the linear slope m and constant c, respectively. Odd functions exhibiting the sign-flip symmetry have $c = 0$. **Scaling** denotes the positive-scaling symmetry of positively 1-homogeneous activations.

Activation	$\sigma(z)$	**Even+Lin**	**Const+Odd**	**Scaling**
`celu`	$\begin{cases}\alpha(\mathrm{e}^{z/\alpha}-1), & z<0\\ z, & z\geq 0\end{cases}$	–	–	✗
`elu`	$\begin{cases}\alpha(\mathrm{e}^{z}-1), & z<0\\ z, & z\geq 0\end{cases}$	–	–	✗
`gelu`	$z\,\Phi(z)$	$1/2$	–	✗
`hard_sigmoid`	$\mathrm{relu6}(z+3)/6$	–	$1/2$	✗
`hard_silu`	$z\,\mathrm{hard_sigmoid}(z)$	$1/2$	–	✗
`hard_tanh`	$\begin{cases}-1, & z\leq -1\\ z, & -1<z<1\\ 1, & z\geq 1\end{cases}$	–	0	✗
`identity`	z	1	0	✓
`leaky_relu`	$\begin{cases}\alpha z, & z<0\\ z, & z\geq 0\end{cases}$	$\frac{1+\alpha}{2}$	–	✓
`log_sigmoid`	$-\log(1+\mathrm{e}^{-z})$	$1/2$	–	✗
`mish`	$z\tanh(\mathrm{softplus}(z))$	–	–	✗
`relu`	$\max(0,z)$	$1/2$	–	✓
`relu6`	$\min(\max(0,z),6)$	–	–	✗
`selu`	$\lambda\begin{cases}\alpha(\mathrm{e}^{z}-1), & z<0\\ z, & z\geq 0\end{cases}$	–	–	✗
`sigmoid`	$(1+\mathrm{e}^{-z})^{-1}$	–	$1/2$	✗
`silu`	$z\,\mathrm{sigmoid}(z)$	$1/2$	–	✗
`soft_sign`	$z/(\lvert z\rvert+1)$	–	0	✗
`softplus`	$\log(1+\mathrm{e}^{z})$	$1/2$	–	✗
`sparse_plus`	$\begin{cases}0, & z\leq -1\\ \frac{1}{4}(z+1)^2, & -1<z<1\\ z, & z\geq 1\end{cases}$	$1/2$	–	✗
`sparse_sigmoid`	$\begin{cases}0, & z\leq -1\\ \frac{1}{2}(z+1), & -1<z<1\\ 1, & z\geq 1\end{cases}$	–	$1/2$	✗
`squareplus`	$\frac{1}{2}(z+\sqrt{z^2+b})$	$1/2$	–	✗
`tanh`	$\tanh(z)$	–	0	✗

Leaky ReLU. From the definition of leaky ReLU, we have

$$\sigma(z) = \begin{cases} \alpha z, & z < 0, \\ z, & z \geq 0 \end{cases} \qquad \text{and} \qquad \sigma(-z) = \begin{cases} -z, & z < 0, \\ -\alpha z, & z \geq 0. \end{cases} \tag{91}$$

For $z \geq 0$, the even component is not constant,

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{z - \alpha z}{2} = \frac{1 - \alpha}{2} z, \tag{92}$$

so leaky ReLU is not constant-odd. The odd component is linear,

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{z + \alpha z}{2} = \frac{1 + \alpha}{2} z, \tag{93}$$

so leaky ReLU is even-linear with $m = \frac{1+\alpha}{2}$.

Log-sigmoid. Using standard logarithm identities, the even component can be written as

$$\begin{aligned} e(z) &= \frac{\sigma(z) + \sigma(-z)}{2} = \frac{-\log(1 + \mathrm{e}^{-z}) - \log(1 + \mathrm{e}^{z})}{2} \\ &= -\frac{\log((1 + \mathrm{e}^{-z})(1 + \mathrm{e}^{z}))}{2} = -\frac{\log(2(1 + \cosh(z)))}{2}, \end{aligned} \tag{94}$$

which is not constant, so the log-sigmoid activation function is not constant-odd. Similarly, for the odd component:

$$\begin{aligned} o(z) &= \frac{\sigma(z) - \sigma(-z)}{2} = \frac{-\log(1 + \mathrm{e}^{-z}) + \log(1 + \mathrm{e}^{z})}{2} \\ &= \frac{1}{2} \log\left(\frac{1 + \mathrm{e}^{z}}{1 + \mathrm{e}^{-z}} \right) = \frac{1}{2} \log\left(\frac{\mathrm{e}^{z}(\mathrm{e}^{-z} + 1)}{1 + \mathrm{e}^{-z}} \right) = \frac{\log(\mathrm{e}^{z})}{2} = \frac{z}{2}. \end{aligned} \tag{95}$$

Thus, the log-sigmoid is even-linear with $m = \frac{1}{2}$.

ReLU. For $z \geq 0$, we have $\max(0, z) = z$ and $\max(0, -z) = 0$. Conversely, $z < 0$ gives $\max(0, z) = 0$ and $\max(0, -z) = -z$. Therefore,

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{1}{2} \begin{cases} z, & z \geq 0, \\ -z, & z < 0 \end{cases} = \frac{|z|}{2}, \tag{96}$$

which is not constant, so ReLU is not constant-odd. Similarly, the odd component is linear,

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{z}{2}, \tag{97}$$

so that ReLU is even-linear with $m = \frac{1}{2}$.

SiLU. By definition, $\mathrm{silu}(z) = z \,\mathrm{sigmoid}(z)$. Since the sigmoid is constant-odd with $c = \frac{1}{2}$, we have $\mathrm{sigmoid}(z) = \frac{1}{2} + o(z)$ for the odd function $o(z) = \mathrm{sigmoid}(z) - \frac{1}{2}$, and thus SiLU satisfies

$$\mathrm{silu}(z) = z\, o(z) + \frac{z}{2}, \tag{98}$$

where the product $z\, o(z)$ is even and $\frac{z}{2}$ is odd. The product $z\, o(z)$ is not constant,

$$z\, o(z) = z \ \mathrm{sigmoid}(z) - \frac{z}{2} = \frac{z}{1 + \mathrm{e}^{-z}} - \frac{z}{2}, \tag{99}$$

so SiLU is not constant-odd. Since the odd component $\frac{z}{2}$ of the sigmoid linear unit is linear, SiLU is even-linear with $m = \frac{1}{2}$.

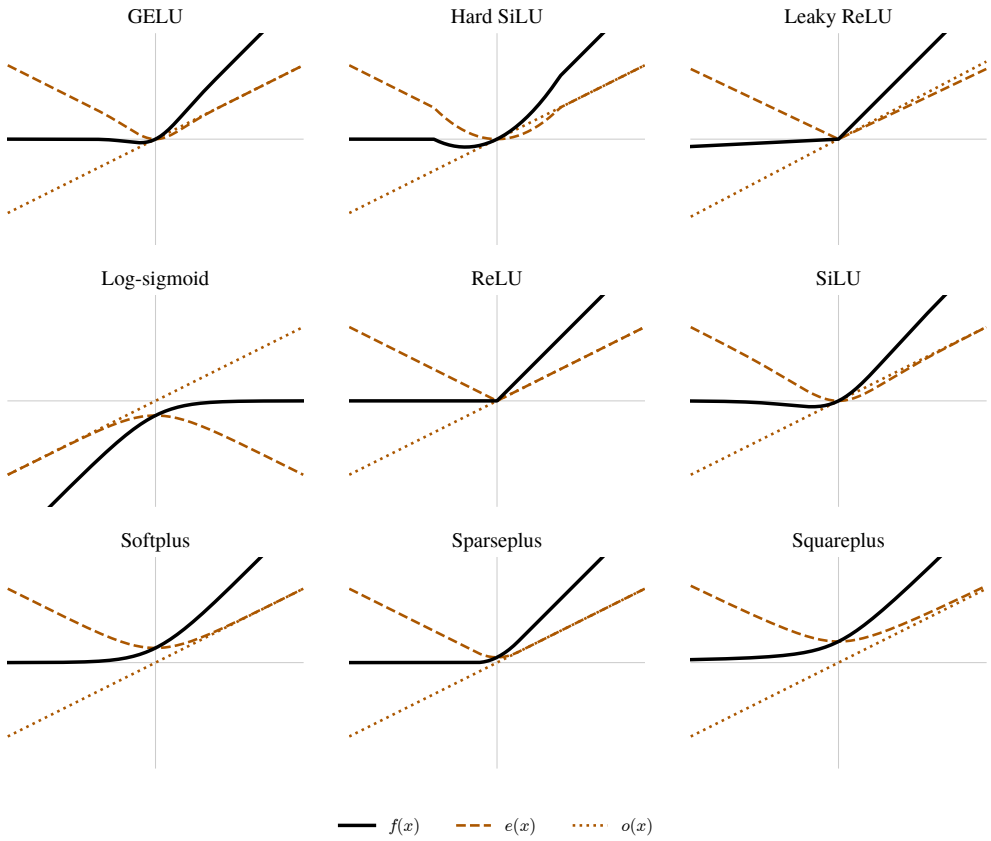

Figure C.1: **Even-linear activations**. Activations from `jax.nn` classified as even-linear, shown with their decomposition into even and odd components. Solid dark lines show the activations, dashed lines their even components, and dotted lines their odd components. In each case, the odd component is a line passing through the origin.

Softplus. Using the identity

$$\sigma(-z) = \log(1 + \mathrm{e}^{-z}) = \log(1 + \mathrm{e}^{z}) - z = \sigma(z) - z \tag{100}$$

of the softplus activation σ, we get

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \sigma(z) - \frac{z}{2}, \tag{101}$$

which is not constant, so softplus is not constant-odd. Using the same identity, the odd component simplifies to

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{z}{2}, \tag{102}$$

which is linear, so softplus is even-linear with $m = \frac{1}{2}$.

Sparseplus. From the definition of the sparseplus activation function, we have

$$\sigma(z) = \begin{cases} 0, & z \le -1, \\ \frac{1}{4}(z+1)^2, & -1 < z < 1, \\ z, & z \ge 1 \end{cases} \quad \text{and} \quad \sigma(-z) = \begin{cases} -z, & z \le -1, \\ \frac{1}{4}(-z+1)^2, & -1 < z < 1, \\ 0, & z \ge 1. \end{cases} \tag{103}$$

For $|z| < 1$, we have

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{(z+1)^2 + (-z+1)^2}{8} = \frac{z^2 + 1}{4} \tag{104}$$

and

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{(z+1)^2 - (-z+1)^2}{8} = \frac{z}{2}. \tag{105}$$

Thus, the even component is not constant, and sparseplus is not constant-odd. The identity $o(z) = \frac{z}{2}$ involving the odd component also holds for $|z| \geq 1$, so sparseplus is even-linear with $m = \frac{1}{2}$.

Squareplus. The squareplus activation function satisfies

$$\sigma(-z) = \frac{-z + \sqrt{z^2 + b}}{2}, \tag{106}$$

so the even component equals

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{(z + \sqrt{z^2 + b}) + (-z + \sqrt{z^2 + b})}{4} = \frac{\sqrt{z^2 + b}}{2}, \tag{107}$$

which is not constant. Thus, squareplus is not constant-odd. The odd component is linear,

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{(z + \sqrt{z^2 + b}) - (-z + \sqrt{z^2 + b})}{4} = \frac{z}{2}, \tag{108}$$

so squareplus is even-linear with $m = \frac{1}{2}$.

C.2 Constant-odd activations

We verify that each of the following `jax.nn` activations is constant-odd in the sense of Definition B.6, and we report the value c of its constant even component.

Hard sigmoid. For $\sigma(z) = \text{relu6}(z+3)/6$, where $\text{relu6}(y) = \min(\max(0, y), 6)$, we have

$$\sigma(z) = \begin{cases} 0, & z \leq -3, \\ (z+3)/6, & -3 < z < 3, \\ 1, & z \geq 3 \end{cases} \quad \text{and} \quad \sigma(-z) = \begin{cases} 1, & z \leq -3, \\ (-z+3)/6, & -3 < z < 3, \\ 0, & z \geq 3. \end{cases} \tag{109}$$

From this, it follows that

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{(z+3) + (-z+3)}{12} = \frac{1}{2}, \qquad |z| < 3. \tag{110}$$

By inspection, the same is true for $|z| \geq 3$. Thus, the hard sigmoid activation function is constant-odd with $c = \frac{1}{2}$. The odd component satisfies $o(z) = -\frac{1}{2}$ for $z \leq -3$. Being constant but nonzero on an unbounded interval precludes the odd component from being linear, so the hard sigmoid is not even-linear.

Hard tanh. Evidently, the hard tanh activation function is odd, making it constant-odd with $c = 0$. Since the hard tanh is piecewise linear but not linear, it is not even-linear.

Identity. The identity $\sigma(z) = z$ is odd, so it is trivially constant-odd with constant $c = 0$.

Sigmoid. The sigmoid $\sigma(z) = (1 + \mathrm{e}^{-z})^{-1}$ satisfies the identity

$$\sigma(-z) = \frac{1}{1 + \mathrm{e}^{z}} = \frac{\mathrm{e}^{-z}}{1 + \mathrm{e}^{-z}} = 1 - \frac{1}{1 + \mathrm{e}^{-z}} = 1 - \sigma(z). \tag{111}$$

Hence, the even component is constant,

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{\sigma(z) + (1 - \sigma(z))}{2} = \frac{1}{2}, \tag{112}$$

so the sigmoid activation function is constant-odd with $c = \frac{1}{2}$. Applying Equation (111) once more shows that the odd component is not linear,

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{\sigma(z) - (1 - \sigma(z))}{2} = \sigma(z) - \frac{1}{2}, \tag{113}$$

so the sigmoid is not even-linear.

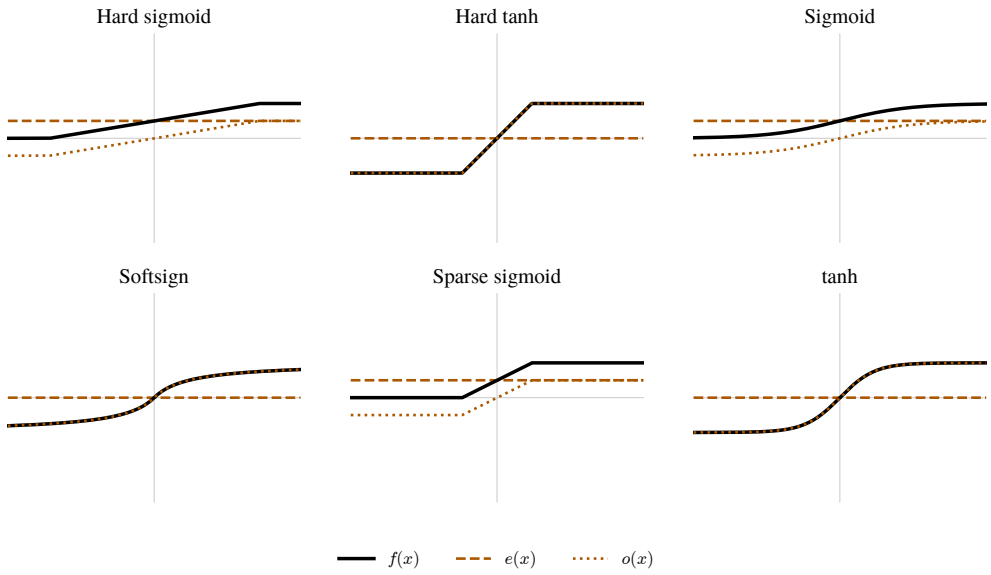

Figure C.2: **Constant-odd activations**. Activations from `jax.nn` classified as constant-odd, shown with their decomposition into even and odd components. Solid dark lines show the activations, dashed lines their even components, and dotted lines their odd components. In each case, the even component is a horizontal line—at $y = 0$ for purely odd functions, where the odd component coincides with the activation itself.

Softsign. The softsign activation function satisfies

$$\sigma(-z) = \frac{-z}{|-z|+1} = -\frac{z}{|z|+1} = -\sigma(z), \tag{114}$$

and thus is odd, making it constant-odd with $c = 0$. Since softsign is not itself linear, it is not even-linear.

Sparse sigmoid. From the definition of the sparse sigmoid, we have

$$\sigma(z) = \begin{cases} 0, & z \le -1, \\ \frac{1}{2}(z+1), & -1 < z < 1, \\ 1, & z \ge 1 \end{cases} \quad \text{and} \quad \sigma(-z) = \begin{cases} 1, & z \le -1, \\ \frac{1}{2}(-z+1), & -1 < z < 1, \\ 0, & z \ge 1. \end{cases} \tag{115}$$

For $|z| < 1$, this gives

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{(z+1) + (-z+1)}{4} = \frac{1}{2}. \tag{116}$$

The same is trivially true for $|z| \ge 1$, so the sparse sigmoid is constant-odd with $c = \frac{1}{2}$. The odd component

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{1}{2}, \qquad z \ge 1, \tag{117}$$

is constant and nonzero on the half-line $[1, \infty)$, and thus cannot be linear, so the sparse sigmoid is not even-linear.

Tanh. The even component of the hyperbolic tangent

$$\tanh(z) = \frac{\mathrm{e}^z - \mathrm{e}^{-z}}{\mathrm{e}^z + \mathrm{e}^{-z}} \tag{118}$$

vanishes:

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{(\mathrm{e}^z - \mathrm{e}^{-z}) + (\mathrm{e}^{-z} - \mathrm{e}^z)}{2(\mathrm{e}^z + \mathrm{e}^{-z})} = 0. \tag{119}$$

Hence, tanh is odd, making it constant-odd with $c = 0$. Since the hyperbolic tangent is not itself linear, it is not even-linear.

C.3 Positively 1-homogeneous activations

We now identify which `jax.nn` activations are positively homogeneous of degree 1. Recall that a function $\sigma\colon \mathbb{R} \to \mathbb{R}$ is positively homogeneous of degree 1 if and only if $\sigma(\alpha z) = \alpha\sigma(z)$ for all $\alpha > 0$ and all $z \in \mathbb{R}$. The following characterization is standard:

Lemma C.1 (Characterization of positively 1-homogeneous functions)**.** *A function $\sigma\colon \mathbb{R} \to \mathbb{R}$ is positively homogeneous of degree 1 if and only if there exist constants $\lambda_-, \lambda_+ \in \mathbb{R}$ such that*

$$\sigma(z) = \begin{cases} \lambda_- z, & z < 0, \\ \lambda_+ z, & z \geq 0. \end{cases} \tag{120}$$

Proof. Any function of the form presented in Equation (120) is readily verified to be positively homogeneous of degree 1. Conversely, suppose σ is positively homogeneous of degree 1. For $z < 0$, we have $-z > 0$ and hence

$$\sigma(z) = -z\,\sigma(-1) = \lambda_- z, \qquad z < 0, \tag{121}$$

with $\lambda_- := -\sigma(-1)$. Similarly, for $z > 0$:

$$\sigma(z) = z\,\sigma(1) = \lambda_+ z, \qquad z > 0, \tag{122}$$

with $\lambda_+ := \sigma(1)$. Finally, positive 1-homogeneity implies $\sigma(0) = \sigma(\alpha \cdot 0) = \alpha\,\sigma(0)$ for every $\alpha > 0$, which requires $\sigma(0) = 0 = \lambda_+ \cdot 0$. □

This characterization immediately implies that positively 1-homogeneous activations form a subclass of even-linear activations.

Corollary C.2. *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ be a positively 1-homogeneous function as in Equation (120). Then, σ is even-linear with even and odd components*

$$e(z) = \frac{\lambda_+ - \lambda_-}{2}\,|z|, \qquad o(z) = \frac{\lambda_+ + \lambda_-}{2}\,z. \tag{123}$$

In particular, σ is nonaffine if and only if $\lambda_+ \neq \lambda_-$.

Proof. For $z < 0$, Equation (120) gives

$$e(z) = \frac{\lambda_- z + \lambda_+(-z)}{2} = \frac{\lambda_+ - \lambda_-}{2}\,|z| \qquad \text{and} \qquad o(z) = \frac{\lambda_- z - \lambda_+(-z)}{2} = \frac{\lambda_+ + \lambda_-}{2}\,z. \tag{124}$$

Similarly, for $z \geq 0$, we have

$$e(z) = \frac{\lambda_+ z + \lambda_-(-z)}{2} = \frac{\lambda_+ - \lambda_-}{2}\,|z| \qquad \text{and} \qquad o(z) = \frac{\lambda_+ z - \lambda_-(-z)}{2} = \frac{\lambda_+ + \lambda_-}{2}\,z. \tag{125}$$

Hence, σ is even-linear in the sense of Definition B.5 with $m = {}^{(\lambda_+ + \lambda_-)}/_2$. The final claim follows directly from the fact that the even component of an affine function $z \mapsto mz + c$ is constant. □

Lemma C.1 also yields the following result:

Corollary C.3 (Unboundedness of positively 1-homogeneous functions)**.** *On each of the half-lines $(-\infty, 0]$ and $[0, \infty)$, a positively homogeneous function of degree 1 is either identically zero or unbounded.*

Corollary C.3 rules out all activation functions except leaky ReLU, ReLU, and the identity from being positively homogeneous of degree 1. These three satisfy the characterization from Lemma C.1, and hence are positively homogeneous of degree 1.

C.4 Activations with neither symmetry property

The remaining `jax.nn` activations are neither even-linear nor constant-odd. We verify this explicitly by computing the even–odd decomposition in each case.

CELU. From the definition of CELU, we have

$$\sigma(z) = \begin{cases} \alpha(\mathrm{e}^{z/\alpha} - 1), & z < 0, \\ z, & z \geq 0 \end{cases} \qquad \text{and} \qquad \sigma(-z) = \begin{cases} -z, & z < 0, \\ \alpha(\mathrm{e}^{-z/\alpha} - 1), & z \geq 0. \end{cases} \tag{126}$$

For $z \geq 0$, this yields

$$e(z) = \frac{z + \alpha(\mathrm{e}^{-z/\alpha} - 1)}{2} \qquad \text{and} \qquad o(z) = \frac{z - \alpha(\mathrm{e}^{-z/\alpha} - 1)}{2}. \tag{127}$$

The linear term in e implies that e is not constant, and the exponential term in o implies that o is not linear. Consequently, CELU is neither constant-odd nor even-linear.

ELU. From the definition of ELU, we have

$$\sigma(z) = \begin{cases} \alpha(\mathrm{e}^{z} - 1), & z < 0, \\ z, & z \geq 0 \end{cases} \qquad \text{and} \qquad \sigma(-z) = \begin{cases} -z, & z < 0, \\ \alpha(\mathrm{e}^{-z} - 1), & z \geq 0. \end{cases} \tag{128}$$

For $z \geq 0$, this yields

$$e(z) = \frac{z + \alpha(\mathrm{e}^{-z} - 1)}{2} \qquad \text{and} \qquad o(z) = \frac{z - \alpha(\mathrm{e}^{-z} - 1)}{2}. \tag{129}$$

As with CELU, the linear term in e implies that e is not constant, and the exponential term in o implies that o is not linear. Again, ELU is neither constant-odd nor even-linear.

Mish. For brevity, write $s(z) \coloneqq \mathrm{softplus}(z) = \log(1 + \mathrm{e}^{z})$. For $\sigma(z) = z \tanh(s(z))$, the even and odd components are

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{z}{2}\big(\tanh(s(z)) - \tanh(s(-z))\big), \tag{130}$$

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = \frac{z}{2}\big(\tanh(s(z)) + \tanh(s(-z))\big). \tag{131}$$

As $z \to \infty$, we have $s(z) \to \infty$ and $s(-z) \to 0$. Therefore,

$$\lim_{z\to\infty} \frac{e(z)}{z} = \frac{1}{2}(1 - 0) = \frac{1}{2}, \qquad \lim_{z\to\infty} \frac{o(z)}{z} = \frac{1}{2}(1 + 0) = \frac{1}{2}. \tag{132}$$

Thus, the even component grows asymptotically as $z/2$ and is not constant. On the other hand, as $z \to 0$, both $s(z)$ and $s(-z)$ converge to $\log 2$, yielding

$$\lim_{z\to 0} \frac{o(z)}{z} = \tanh(\log 2) = \frac{\mathrm{e}^{2\log 2} - 1}{\mathrm{e}^{2\log 2} + 1} = \frac{\mathrm{e}^{\log 2^2} - 1}{\mathrm{e}^{\log 2^2} + 1} = \frac{3}{5}. \tag{133}$$

Since $o(z)/z$ has different limits as $z \to 0$ and $z \to \infty$, it cannot be constant, and hence the odd component is not linear. Consequently, Mish is neither constant-odd nor even-linear.

ReLU6. From the definition of ReLU6, we have

$$\sigma(z) = \begin{cases} 0, & z \leq 0, \\ z, & 0 < z < 6, \\ 6, & z \geq 6 \end{cases} \qquad \text{and} \qquad \sigma(-z) = \begin{cases} 6, & z \leq -6, \\ -z, & -6 < z < 0, \\ 0, & z \geq 0. \end{cases} \tag{134}$$

For $0 < z < 6$, the even component is given by

$$e(z) = \frac{\sigma(z) + \sigma(-z)}{2} = \frac{z}{2}, \tag{135}$$

which is not constant, so ReLU6 is not constant-odd. At the same time, the odd component satisfies

$$o(z) = \frac{\sigma(z) - \sigma(-z)}{2} = 3, \qquad z \geq 6. \tag{136}$$

Being constant and nonzero on an unbounded interval implies that the odd component is not linear, so ReLU6 is not even-linear.

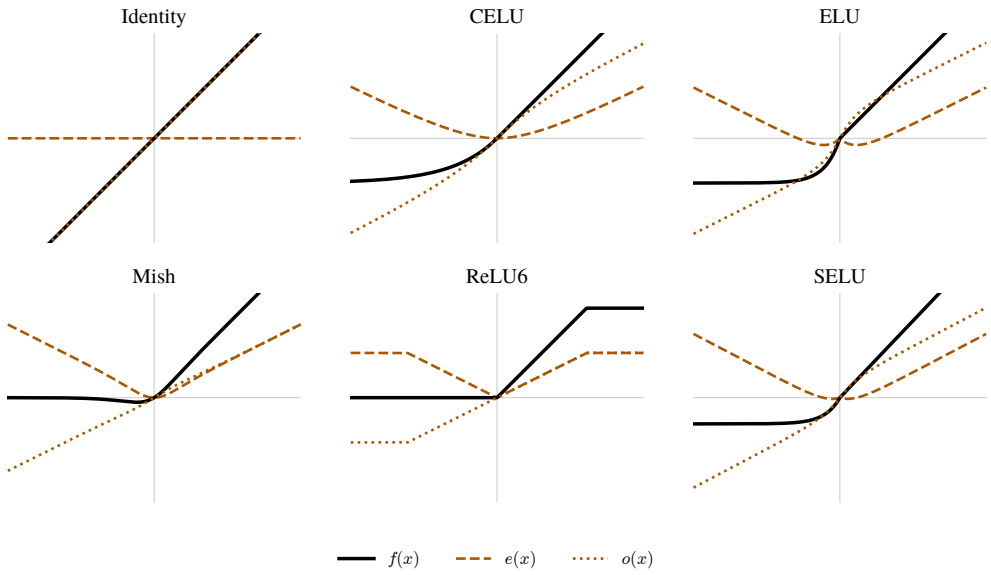

Figure C.3: **Activations with both symmetry properties or neither**. Activations from `jax.nn` that are either both even-linear and constant-odd (the identity) or neither (the asymmetric activations), shown with their decomposition into even and odd components. Solid dark lines show the activations, dashed lines their even components, and dotted lines their odd components.

SELU. From the definition of SELU, we have

$$\sigma(z) = \lambda \begin{cases} \alpha(\mathrm{e}^z - 1), & z < 0, \\ z, & z \geq 0 \end{cases} \quad \text{and} \quad \sigma(-z) = \lambda \begin{cases} -z, & z < 0, \\ \alpha(\mathrm{e}^{-z} - 1), & z \geq 0. \end{cases} \tag{137}$$

For $z \geq 0$, this yields

$$e(z) = \lambda \frac{z + \alpha(\mathrm{e}^{-z} - 1)}{2} \quad \text{and} \quad o(z) = \lambda \frac{z - \alpha(\mathrm{e}^{-z} - 1)}{2}. \tag{138}$$

The even component is clearly not constant, so SELU is not constant-odd. At the same time, the odd component is not linear due to the presence of the exponential term, so SELU is not even-linear either.

D Feature transformations within symmetry orbits

This appendix provides proofs and extensions of the results in Section 4. We first introduce duplication patterns and duplication matrices and establish identities governing their action on vectors and matrices (Appendix D.1). We then show that every parameter symmetry cataloged in Appendices B.1 to B.3 induces a finite composition of feature additions, duplications, scalings, and their inverses (Appendix D.2). Next, we extend the characterization of hidden-activation matrices within a symmetry orbit in Proposition 4.2 to all activation classes considered in this work (Appendix D.3). Finally, under the activation assumptions of Corollary 4.3, we characterize irreducible parameterizations within a symmetry orbit, establish their uniqueness up to generic reparameterization, and prove that their feature rows persist up to nonzero scaling throughout the orbit (Appendix D.4).

D.1 Duplication matrices

To express feature duplication as matrix multiplication, we first record the number of copies of each feature row in a *duplication pattern* and then construct the corresponding *duplication matrix*.

Definition D.1 (Duplication pattern)**.** Let $L \in \mathbb{N}$ denote the number of feature rows before duplication. A *duplication pattern* is a vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)^\top$, where $\nu_i \in \mathbb{N}_{>0}$ denotes the number of copies

Table D.1: Effect of multiplying vectors and matrices by the duplication matrix $\mathbf{D}_{\boldsymbol{\nu}} \in \{0,1\}^{N_h \times L}$ and its transpose.

Input	Transformation	Result
$\mathbf{p} \in \mathbb{R}^L$	$\mathbf{D}_{\boldsymbol{\nu}}\mathbf{p} \in \mathbb{R}^{N_h}$	repeats the ith entry of $\mathbf{p}$ exactly ν_i times
$\mathbf{M} \in \mathbb{R}^{L \times Q}$	$\mathbf{D}_{\boldsymbol{\nu}}\mathbf{M} \in \mathbb{R}^{N_h \times Q}$	repeats each row i of $\mathbf{M}$ exactly ν_i times
$\mathbf{N} \in \mathbb{R}^{Q \times L}$	$\mathbf{N}\mathbf{D}_{\boldsymbol{\nu}}^\top \in \mathbb{R}^{Q \times N_h}$	repeats each column i of $\mathbf{N}$ exactly ν_i times

of the ith row after duplication. We refer to the resulting number of rows,

$$N_h = \sum_{i=1}^{L} \nu_i, \tag{139}$$

as the duplication pattern's *induced width*.

Definition D.2 (Duplication matrix)**.** Given a duplication pattern $\boldsymbol{\nu} \in \mathbb{N}_{>0}^L$ with induced width N_h, define the *duplication matrix* $\mathbf{D}_{\boldsymbol{\nu}} \in \{0,1\}^{N_h \times L}$ by

$$\mathbf{D}_{\boldsymbol{\nu}} \coloneqq \begin{bmatrix} \mathbf{1}_{\nu_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\nu_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{\nu_L} \end{bmatrix}, \tag{140}$$

where each $\mathbf{1}_{\nu_i} \in \mathbb{R}^{\nu_i}$ is a vector of all ones. Equivalently,

$$(\mathbf{D}_{\boldsymbol{\nu}})_{rj} \coloneqq \begin{cases} 1, & \text{if } \sum_{i=1}^{j-1} \nu_i < r \le \sum_{i=1}^{j} \nu_i, \\ 0, & \text{otherwise.} \end{cases} \tag{141}$$

Remark D.3 (Index shorthand). For a vector $\boldsymbol{\nu} \in \mathbb{R}^L$, we denote the sum of its first $j-1$ entries by

$$\nu_{<j} \coloneqq \sum_{i=1}^{j-1} \nu_i, \qquad j = 1, \ldots, L+1, \tag{142}$$

with $\nu_{<1} \coloneqq 0$. With this shorthand notation, we can rewrite the definition of the duplication matrix $\mathbf{D}_{\boldsymbol{\nu}}$ given in Equation (141) as

$$(\mathbf{D}_{\boldsymbol{\nu}})_{rj} = \begin{cases} 1, & \text{if } \nu_{<j} < r \le \nu_{<(j+1)}, \\ 0, & \text{otherwise.} \end{cases} \tag{143}$$

Left multiplication by $\mathbf{D}_{\boldsymbol{\nu}}$ repeats entries of a column vector or rows of a matrix according to the duplication pattern $\boldsymbol{\nu}$; right multiplication by $\mathbf{D}_{\boldsymbol{\nu}}^\top$ repeats columns analogously. These operations are summarized in Table D.1 and made precise by the following result.

Lemma D.4 (Multiplication by duplication matrices)**.** *Let $\boldsymbol{\nu} \in \mathbb{N}_{>0}^L$ be a duplication pattern with induced width N_h. Further, let $\mathbf{p} \in \mathbb{R}^L$, let $\mathbf{M} \in \mathbb{R}^{L \times Q}$, and let $\mathbf{N} \in \mathbb{R}^{Q \times L}$. Then, whenever $\nu_{<j} < r \le \nu_{<(j+1)}$, the following identities hold:*

(i) $(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{p})_r = p_j$,

(ii) $(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{M})_{r,:} = \mathbf{M}_{j,:}$,

(iii) $(\mathbf{N}\mathbf{D}_{\boldsymbol{\nu}}^\top)_{:,r} = \mathbf{N}_{:,j}$.

Proof. For part (i), we have

$$(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{p})_r = \sum_{i=1}^{L} (\mathbf{D}_{\boldsymbol{\nu}})_{ri} p_i. \tag{144}$$

Since the intervals $(\nu_{<i}, \nu_{<(i+1)}]$ partition $(0, N_h]$, each $r \in \{1, \ldots, N_h\}$ lies in exactly one such interval, say for index j. Then, $(\mathbf{D}_{\boldsymbol{\nu}})_{rj} = 1$ and $(\mathbf{D}_{\boldsymbol{\nu}})_{ri} = 0$ for $i \neq j$, giving $(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{p})_r = p_j$. Part (ii) follows by applying (i) to each column of $\mathbf{M}$, and part (iii) follows by transposing (ii). □

Table D.2: Decomposition of parameter symmetries into primitive feature transformations. Checkmarks indicate the primitives appearing in the displayed decompositions, possibly acting trivially when duplication counts equal 1. For width-changing operations, the table describes the addition or replacement direction; reverse operations use the corresponding inverse primitives.

Symmetry	Addition	Duplication	Scaling
Permutation	✗	✗	✗
Positive scaling	✗	✗	✓
Sign flip	✗	✗	✓
Zero-neuron group	✓	✓	✗
Duplicate-neuron group	✗	✓	✗
Constant neuron	✓	✗	✗
Linear-neuron group	✓	✓	✗
Linear-duplicate-neuron group	✓	✓	✗
Constant-neuron group	✓	✓	✗
Constant-duplicate-neuron group	✓	✓	✗

The next identity describes how weights assigned to individual copies combine when those copies are grouped by their original row. Each diagonal entry of the resulting weighted Gram matrix is the sum of the weights over one duplication group.

Lemma D.5 (Weighted Gram matrix of a duplication matrix)**.** *Let $\boldsymbol{\nu} \in \mathbb{N}_{>0}^{L}$ be a duplication pattern with induced width N_h, and let $\boldsymbol{\omega} \in \mathbb{R}^{N_h}$. Then, the $\operatorname{diag}(\boldsymbol{\omega})$-weighted Gram matrix of the duplication matrix $\mathbf{D}_{\boldsymbol{\nu}}$ is the diagonal matrix given by*

$$\mathbf{D}_{\boldsymbol{\nu}}^{\top} \operatorname{diag}(\boldsymbol{\omega}) \mathbf{D}_{\boldsymbol{\nu}} = \operatorname{diag}(\mathbf{D}_{\boldsymbol{\nu}}^{\top} \boldsymbol{\omega}). \tag{145}$$

Proof. Let $\mathbf{d}_j$ denote the jth column of $\mathbf{D}_{\boldsymbol{\nu}}$. The entry at position (r, j) of the weighted Gram matrix $\mathbf{D}_{\boldsymbol{\nu}}^{\top} \operatorname{diag}(\boldsymbol{\omega}) \mathbf{D}_{\boldsymbol{\nu}}$ equals

$$(\mathbf{D}_{\boldsymbol{\nu}}^{\top} \operatorname{diag}(\boldsymbol{\omega}) \mathbf{D}_{\boldsymbol{\nu}})_{rj} = \mathbf{d}_r^{\top} \operatorname{diag}(\boldsymbol{\omega}) \mathbf{d}_j = \sum_{i=1}^{N_h} \omega_i (\mathbf{d}_r)_i (\mathbf{d}_j)_i. \tag{146}$$

This sum accumulates the weights ω_i over all rows i in which columns r and j of $\mathbf{D}_{\boldsymbol{\nu}}$ are both equal to 1. Since each row of $\mathbf{D}_{\boldsymbol{\nu}}$ contains *exactly one* nonzero entry, no two distinct columns can have ones in the same row, i.e., $(\mathbf{d}_r)_i (\mathbf{d}_j)_i = 0$ whenever $r \neq j$. For the diagonal entries $r = j$, each entry of $\mathbf{d}_j$ is either 0 or 1, so $(\mathbf{d}_j)_i^2 = (\mathbf{d}_j)_i$, and hence

$$(\mathbf{D}_{\boldsymbol{\nu}}^{\top} \operatorname{diag}(\boldsymbol{\omega}) \mathbf{D}_{\boldsymbol{\nu}})_{jj} = \sum_{i=1}^{N_h} \omega_i (\mathbf{d}_j)_i^2 = \sum_{i=1}^{N_h} \omega_i (\mathbf{d}_j)_i = \mathbf{d}_j^{\top} \boldsymbol{\omega} = (\mathbf{D}_{\boldsymbol{\nu}}^{\top} \boldsymbol{\omega})_j, \tag{147}$$

completing the proof. □

D.2 Feature-level characterization of parameter symmetries

We now prove Proposition 4.1 by translating the cataloged parameter symmetries into operations on feature rows. The decompositions established in the proof are summarized in Table D.2.

Proof of Proposition 4.1. We verify the claim for the parameter symmetries cataloged in Appendices B.1 to B.3. Throughout this proof, $\mathbf{H}$ denotes the hidden-activation matrix before the transformation, and N_h its number of rows.

Permutation symmetry. Permuting hidden neurons solely permutes the rows of the hidden-activation matrix $\mathbf{H}$. Because Proposition 4.1 is stated up to row permutations of $\mathbf{H}$, we may choose any convenient row ordering throughout this proof. We place newly introduced rows after existing ones and duplicate copies of a given row consecutively.

Positive-scaling symmetry. For positively homogeneous activations of degree 1, rescaling the jth neuron's incoming weights $\mathbf{w}_j$ and hidden bias b_j by $\alpha_j > 0$, and inversely scaling its readout weights $\mathbf{a}_j$ by α_j^{-1}, multiplies the jth row of $\mathbf{H}$ by α_j. This is an instance of feature scaling:

$$\mathbf{H} \mapsto \operatorname{diag}(\boldsymbol{\alpha})\mathbf{H}, \qquad \alpha_i = \begin{cases} 1, & i \neq j, \\ \alpha_j, & i = j. \end{cases} \tag{148}$$

Sign-flip symmetry. For odd activations, flipping the signs of the jth neuron's incoming weights $\mathbf{w}_j$, hidden bias b_j, and readout weights $\mathbf{a}_j$ leaves the realized function unchanged. Since $\sigma(-z) = -\sigma(z)$, the feature computed by the jth neuron changes sign. This is another instance of feature scaling:

$$\mathbf{H} \mapsto \operatorname{diag}(\boldsymbol{\alpha})\mathbf{H}, \qquad \alpha_i = \begin{cases} 1, & i \neq j, \\ -1, & i = j. \end{cases} \tag{149}$$

For even activations, by contrast, flipping the incoming weights and hidden bias leaves the feature unchanged because $\sigma(-z) = \sigma(z)$. Thus, the sign flip is absorbed by the activation, so $\mathbf{H}$ is unchanged, corresponding to feature scaling by the identity.

Zero-neuron groups. Since all neurons in a zero-neuron group share incoming parameters $\mathbf{w}$ and b, they induce the same feature vector $\mathbf{u} \coloneqq \sigma(\mathbf{w}^\top \mathbf{X} + b\mathbf{1}^\top) \in \mathbb{R}^{1\times P}$. Introducing a zero-neuron group indexed by $\mathcal{Z}$ therefore amounts to adding this feature once and duplicating it according to the group size:

$$\mathbf{H} \mapsto \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H} \\ \mathbf{u} \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} \mathbf{1}_{N_h} \\ |\mathcal{Z}| \end{bmatrix}. \tag{150}$$

Duplicate-neuron groups. Replacing the jth neuron with a duplicate-neuron group indexed by $\mathcal{D}$ replicates the jth row of $\mathbf{H}$ into $|\mathcal{D}|$ copies. This is an instance of feature duplication:

$$\mathbf{H} \mapsto \mathbf{D}_{\boldsymbol{\nu}}\mathbf{H}, \qquad \nu_i = \begin{cases} 1, & i \neq j, \\ |\mathcal{D}|, & i = j. \end{cases} \tag{151}$$

Constant neurons. Since a constant neuron has vanishing incoming weights, its output $\sigma(b)$ is independent of the input, producing the constant feature vector $\mathbf{u} \coloneqq \sigma(b)\mathbf{1}^\top \in \mathbb{R}^{1\times P}$. Adding a constant neuron is therefore a feature addition of the form

$$\mathbf{H} \mapsto \begin{bmatrix} \mathbf{H} \\ \mathbf{u} \end{bmatrix}. \tag{152}$$

Linear-neuron groups. For an even-linear activation $\sigma(z) = e(z) + mz$, the aligned and opposite subgroups of a linear-neuron group produce the feature rows associated with the incoming parameters $(\mathbf{w}, b)$ and $(-\mathbf{w}, -b)$, respectively. Writing

$$\mathbf{z} \coloneqq \mathbf{w}^\top \mathbf{X} + b\mathbf{1}^\top \in \mathbb{R}^{1\times P} \tag{153}$$

for the pre-activation feature vector, aligned neurons produce $\mathbf{u}^+ \coloneqq e(\mathbf{z}) + m\mathbf{z}$, whereas opposite neurons produce $\mathbf{u}^- \coloneqq e(\mathbf{z}) - m\mathbf{z}$, with e applied element-wise. These rows need not be distinct. Adding a linear-neuron group is therefore a feature addition of these two rows, followed by a feature duplication that replicates $\mathbf{u}^+$ for each aligned neuron and $\mathbf{u}^-$ for each opposite neuron:

$$\mathbf{H} \mapsto \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H} \\ \mathbf{u}^+ \\ \mathbf{u}^- \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} \mathbf{1}_{N_h} \\ |\mathcal{N}^+| \\ |\mathcal{N}^-| \end{bmatrix}. \tag{154}$$

Linear-duplicate-neuron groups. Since aligned neurons share the incoming weights and hidden bias of the jth neuron, they replicate the jth row of $\mathbf{H}$. Opposite neurons have sign-flipped incoming weights and hidden bias. Writing

$$\mathbf{z}_j \coloneqq \mathbf{w}_j^\top \mathbf{X} + b_j\mathbf{1}^\top \in \mathbb{R}^{1\times P} \tag{155}$$

for the pre-activation feature vector of the jth neuron, opposite neurons produce $\mathbf{u}^- \coloneqq e(\mathbf{z}_j) - m\mathbf{z}_j$. Replacing the jth neuron with a linear-duplicate-neuron group is therefore a feature addition of $\mathbf{u}^-$,

followed by a feature duplication that replicates the jth row of $\mathbf{H}$ for each aligned neuron and $\mathbf{u}^-$ for each opposite neuron:

$$\mathbf{H} \mapsto \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H} \\ \mathbf{u}^- \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} \boldsymbol{\nu}^+ \\ |\mathcal{N}^-| \end{bmatrix}, \qquad \nu_i^+ = \begin{cases} 1, & i \neq j, \\ |\mathcal{N}^+|, & i = j. \end{cases} \tag{156}$$

Constant-neuron groups. For a constant-odd activation $\sigma(z) = c + o(z)$, aligned and opposite subgroups produce $\mathbf{u}^+ \coloneqq c\mathbf{1}^\top + o(\mathbf{z})$ and $\mathbf{u}^- \coloneqq c\mathbf{1}^\top - o(\mathbf{z})$, respectively, where $\mathbf{z}$ is defined in Equation (153) and o is applied element-wise. The resulting transformation is otherwise identical to Equation (154).

Constant-duplicate-neuron groups. Opposite neurons produce $\mathbf{u}^- \coloneqq c\mathbf{1}^\top - o(\mathbf{z}_j)$, where $\mathbf{z}_j$ is defined in Equation (155). The resulting transformation is otherwise identical to Equation (156).

Reverse operations. Each displayed transformation can be reversed, on its image, by applying the corresponding inverse primitives in reverse order. In particular, removing an added group first merges duplicated rows and then removes the newly appended rows. Collapsing a duplicate-neuron group only merges its copies, whereas collapsing a linear- or constant-duplicate-neuron group additionally removes the appended opposite feature. Because both aligned and opposite subgroups are nonempty, all duplication counts are strictly positive: inverse duplication therefore retains at least one copy of every pre-duplication row, while rows are removed only through inverse feature addition. Scaling is reversed by reciprocal scaling.

Function-preserving combinations. As established in Appendices B.2.1 and B.3.6, operations involving individual constant neurons or activation-dependent symmetry groups may need to be performed jointly for their residual contributions to cancel. For any such function-preserving combination, the induced transformation of hidden activations is obtained by composing the feature-level transformations of its constituent operations.

Together, these decompositions establish Proposition 4.1. □

D.3 Hidden-activation matrices within a symmetry orbit

The preceding subsection describes how individual parameter symmetries transform hidden-activation matrices. We next prove the main-text characterization of hidden-activation matrices within a symmetry orbit for positively 1-homogeneous activations (Proposition 4.2) using the orbit invariants established in Appendix B.4, and then extend this characterization to all activation classes considered here.

For each activation class, fix a symmetry orbit $\mathcal{O}$ with essential parameter classes $\mathcal{E}$. We collect below the corresponding definitions of the matrix $\mathbf{F}_\mathcal{E}$ introduced in Section 4.2, including the positively 1-homogeneous case already treated in the main text.

Positively homogeneous activations of degree 1. If σ is positively homogeneous of degree 1, each parameter class $q \in \mathcal{E}$ is of the form $q = \{\alpha\overline{\mathbf{w}} \mid \alpha \in \mathbb{R}_{\neq 0}\}$. For each $q \in \mathcal{E}$, choose a unit-norm representative $\overline{\mathbf{w}}_q \in q$, and let $\mathbf{F}_\mathcal{E} \in \mathbb{R}^{2|\mathcal{E}| \times P}$ collect the two features $(\mathbf{f}_q^+)^\top = \sigma(\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}})$ and $(\mathbf{f}_q^-)^\top = \sigma(-\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}})$.

Even-linear and constant-odd activations. If σ is even-linear but not positively homogeneous of degree 1, or is constant-odd, each parameter class q contains exactly two opposite incoming-parameter vectors $\overline{\mathbf{w}}_q$ and $-\overline{\mathbf{w}}_q$. For each $q \in \mathcal{E}$, let $\mathbf{F}_\mathcal{E} \in \mathbb{R}^{2|\mathcal{E}| \times P}$ collect the two features $(\mathbf{f}_q^+)^\top = \sigma(\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}})$ and $(\mathbf{f}_q^-)^\top = \sigma(-\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}})$.

Activations with neither symmetry property. If σ is neither even-linear nor constant-odd, every parameter class q is a singleton $\{\overline{\mathbf{w}}_q\}$. For each $q \in \mathcal{E}$, let $\mathbf{F}_\mathcal{E} \in \mathbb{R}^{|\mathcal{E}| \times P}$ collect the feature $\mathbf{f}_q^\top = \sigma(\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}})$.

We let each incoming-parameter vector $\pm\overline{\mathbf{w}}_q$ specified above contribute a separate row to $\mathbf{F}_\mathcal{E}$, even when the oppositely oriented vectors $\overline{\mathbf{w}}_q$ and $-\overline{\mathbf{w}}_q$ happen to produce identical features $\mathbf{f}_q^+ = \mathbf{f}_q^-$ on the given inputs $\mathbf{X}$, and we arrange these rows in a fixed order.

Proposition D.6 (Hidden-activation matrices within a symmetry orbit)**.** *Fix a symmetry orbit $\mathcal{O}$ with nonaffine activation σ and essential parameter classes $\mathcal{E}$, and let $\mathbf{F}_{\mathcal{E}}$ be defined as above. Every parameterization $\boldsymbol{\theta}$ of width N_h in $\mathcal{O}$ has a hidden-activation matrix of the form*

$$\mathbf{H} = \operatorname{diag}(\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{F}_{\mathcal{E},\mathcal{I}} \\ \mathbf{U} \end{bmatrix}, \qquad \boldsymbol{\alpha} \in \mathbb{R}_{>0}^{N_h}, \quad \boldsymbol{\nu} \in \mathbb{N}_{>0}^{|\mathcal{I}|+K}, \tag{157}$$

for some $K \geq 0$ and $\mathbf{U} \in \mathbb{R}^{K\times P}$, with $\sum_\ell \nu_\ell = N_h$. Here, $\mathbf{F}_{\mathcal{E},\mathcal{I}}$ is the submatrix of $\mathbf{F}_{\mathcal{E}}$ indexed by $\mathcal{I}$. The index set $\mathcal{I}$ selects exactly those rows of $\mathbf{F}_{\mathcal{E}}$ whose associated incoming-parameter vectors are represented by neurons of $\boldsymbol{\theta}$, up to positive scaling when σ is positively homogeneous of degree 1. For every essential class $q \in \mathcal{E}$, at least one of its associated rows is selected. The matrix $\mathbf{U}$ collects pairwise distinct features from nonessential classes and constant neurons. If σ is not positively homogeneous of degree 1, $\boldsymbol{\alpha}$ may be chosen as the all-ones vector.

Proof. Let $\boldsymbol{\theta}$ be any parameterization in $\mathcal{O}$. By Proposition B.15,

$$\boldsymbol{\beta}_q(\boldsymbol{\theta}) = \boldsymbol{\beta}_q \neq \mathbf{0}, \qquad q \in \mathcal{E}. \tag{158}$$

Since the class aggregate of an unrepresented parameter class is zero, every essential class must be represented in $\boldsymbol{\theta}$, that is, $\mathcal{E} \subseteq \mathcal{Q}(\boldsymbol{\theta})$.

Suppose first that σ is positively homogeneous of degree 1. For each neuron $j \in \mathcal{J}_q$, $q \in \mathcal{E}$, positive 1-homogeneity gives

$$\mathbf{H}_{j,:} = \sigma(\overline{\mathbf{w}}_j^\top \overline{\mathbf{X}}) = \alpha_j\, \sigma(\pm\overline{\mathbf{w}}_q^\top \overline{\mathbf{X}}), \qquad \alpha_j \coloneqq \|\overline{\mathbf{w}}_j\| > 0, \tag{159}$$

with the sign determined by whether $j \in \mathcal{J}_q^+$ or $j \in \mathcal{J}_q^-$. Thus, neuron j corresponds to the row of $\mathbf{F}_{\mathcal{E}}$ associated with $\pm\overline{\mathbf{w}}_q$, with scaling factor α_j.

If σ is even-linear but not positively homogeneous of degree 1, or is constant-odd, each neuron in an essential class q has one of the two incoming-parameter vectors $\overline{\mathbf{w}}_q$ and $-\overline{\mathbf{w}}_q$ in q. Its feature therefore coincides exactly with the corresponding row of $\mathbf{F}_{\mathcal{E}}$. We associate each such neuron with that row and set $\alpha_j \coloneqq 1$. The same argument applies when σ is neither even-linear nor constant-odd, except that each essential class contains a single incoming-parameter vector.

Thus, in every activation case, each neuron of $\boldsymbol{\theta}$ belonging to an essential parameter class has been associated with a row of $\mathbf{F}_{\mathcal{E}}$. Let $\mathcal{I}$ index exactly those rows of $\mathbf{F}_{\mathcal{E}}$ that have been associated with at least one such neuron. Since every essential class is represented in $\boldsymbol{\theta}$, the set $\mathcal{I}$ contains at least one index for every $q \in \mathcal{E}$. Ordering the rows of $\mathbf{F}_{\mathcal{E},\mathcal{I}}$ as inherited from $\mathbf{F}_{\mathcal{E}}$, let ν_ℓ denote the number of neurons associated with the ℓth row of $\mathbf{F}_{\mathcal{E},\mathcal{I}}$, for $\ell = 1, \ldots, |\mathcal{I}|$. By construction, $\nu_\ell > 0$.

Every neuron not yet associated with a row of $\mathbf{F}_{\mathcal{E}}$ either belongs to a nonessential parameter class or is constant. Let $\mathbf{U} \in \mathbb{R}^{K\times P}$ collect each of the K *distinct* feature rows generated by these neurons. For each remaining neuron j, associate it with the unique row of $\mathbf{U}$ equal to its feature row, and set $\alpha_j \coloneqq 1$. For $k = 1, \ldots, K$, let $\nu_{|\mathcal{I}|+k}$ denote the number of neurons associated with the kth row of $\mathbf{U}$. Since each row of $\mathbf{U}$ represents at least one neuron, these multiplicities are strictly positive.

Ordering the neurons according to these associations, the definition of the duplication matrix yields Equation (157). Since every neuron in the parameterization $\boldsymbol{\theta}$ is counted exactly once,

$$\sum_{\ell=1}^{|\mathcal{I}|+K} \nu_\ell = N_h. \tag{160}$$

All scaling factors are positive and, when σ is not positively homogeneous of degree 1, equal to 1. Restricting the result to positively homogeneous activations of degree 1 proves Proposition 4.2 from the main text. □

Proposition D.6 gives a necessary structural description of hidden-activation matrices within a fixed symmetry orbit, but not a converse realizability result. In particular, it does *not* guarantee that every choice of $\boldsymbol{\alpha}$, $\boldsymbol{\nu}$, $\mathcal{I}$, and $\mathbf{U}$ is realized by a parameterization in $\mathcal{O}$.

D.4 Persistence of irreducible features

For activations that are neither even-linear nor constant-odd, as well as for purely even and purely odd activations, we prove that every feature computed by an irreducible parameterization must persist throughout its symmetry orbit (Corollary 4.3). We first characterize irreducible parameterizations within an orbit and then describe their remaining parameter freedom. Finally, we return to this characterization to prove that every irreducible feature persists throughout the orbit.

Lemma D.7 (Irreducible parameterizations within a symmetry orbit)**.** *Suppose that σ is neither even-linear nor constant-odd, or is purely even or purely odd. Fix a symmetry orbit $\mathcal{O}$ with essential parameter classes $\mathcal{E}$ and global residual $\mathbf{r}$. A parameterization in $\mathcal{O}$ is irreducible if and only if it contains exactly one neuron from each essential parameter class, no neurons from nonessential parameter classes, and exactly one constant neuron when $\mathbf{r} \neq \mathbf{0}$ and none when $\mathbf{r} = \mathbf{0}$.*

Proof. We first derive a lower bound on the width of any parameterization in $\mathcal{O}$. Under the assumptions of the lemma, nonconstant neurons make no contribution to the global residual $\mathbf{r}$. For activations that are neither even-linear nor constant-odd, this follows directly from the definition of $\mathbf{r}$ in Equation (72). For purely even activations, $m = 0$ in the even-linear decomposition, so the contribution of nonconstant neurons to $\mathbf{r}$ vanishes in Equations (66) and (68), regardless of whether σ is also positively 1-homogeneous. Similarly, for purely odd activations, $c = 0$ in the constant-odd decomposition, so nonconstant neurons make no contribution to $\mathbf{r}$ in Equation (70). Hence, every parameterization $\boldsymbol{\theta}$ in $\mathcal{O}$ satisfies

$$\mathbf{r}(\mathbf{x}) = \sum_{j \in \mathcal{J}_0} \mathbf{a}_j \, \sigma(b_j). \tag{161}$$

Moreover, each essential parameter class $q \in \mathcal{E}$ has nonzero aggregate $\boldsymbol{\beta}_q$ and must therefore be represented by at least one nonconstant neuron. If $\mathbf{r} \neq \mathbf{0}$, the preceding identity additionally requires at least one constant neuron. Hence, the width N_h of any parameterization in $\mathcal{O}$ satisfies

$$N_h \geq \begin{cases} |\mathcal{E}|, & \mathbf{r} = \mathbf{0}, \\ |\mathcal{E}| + 1, & \mathbf{r} \neq \mathbf{0}. \end{cases} \tag{162}$$

We next show that this lower bound is attainable. For each essential class q, construct a single nonconstant neuron with incoming parameters $\overline{\mathbf{w}}_q$, which therefore computes ϕ_q. Assigning this neuron readout $\boldsymbol{\beta}_q$ reproduces exactly the aggregate associated with q (Equations (65), (67), (69), and (71)). If $\mathbf{r} \neq \mathbf{0}$, choose a hidden bias b with $\sigma(b) \neq 0$ and add one constant neuron with readout

$$\mathbf{a} \coloneqq \frac{\mathbf{r}(\mathbf{0})}{\sigma(b)}. \tag{163}$$

The resulting parameterization has aggregate $\boldsymbol{\beta}_q$ for every essential class q, zero aggregate for every nonessential class, and global residual $\mathbf{r}$. By Proposition B.15, it therefore belongs to the fixed orbit $\mathcal{O}$ and attains the lower bound in Equation (162). Since irreducible parameterizations are precisely the minimum-width parameterizations in $\mathcal{O}$, equality in this bound holds exactly when there is one neuron from each essential class, no neurons from nonessential classes, and one constant neuron if and only if $\mathbf{r} \neq \mathbf{0}$. □

The preceding lemma determines the number of neurons in an irreducible parameterization and assigns one neuron to each essential parameter class, with a constant neuron present precisely when the residual is nonzero. We next determine the freedom that remains in each neuron's parameters when the orbit invariants are held fixed.

Corollary D.8 (Uniqueness up to generic reparameterization)**.** *Suppose that σ is neither even-linear nor constant-odd, or is purely even or purely odd. Let $\boldsymbol{\theta}^\star \sim \boldsymbol{\xi}^\star$ be irreducible parameterizations in the same symmetry orbit, with global residual $\mathbf{r}$. Up to a permutation of neurons, corresponding nonconstant neurons differ only by the following transformations:*

(i) If σ is purely even and positively 1-homogeneous,

$$(\overline{\mathbf{w}}, \mathbf{a}) \mapsto (t\overline{\mathbf{w}}, |t|^{-1}\mathbf{a}), \qquad t \in \mathbb{R}_{\neq 0}. \tag{164}$$

(ii) If σ is purely even but not positively 1-homogeneous,

$$(\overline{\mathbf{w}}, \mathbf{a}) \mapsto (\varepsilon\overline{\mathbf{w}}, \mathbf{a}), \qquad \varepsilon \in \{-1, 1\}. \tag{165}$$

(iii) If σ is purely odd,

$$(\overline{\mathbf{w}}, \mathbf{a}) \mapsto (\varepsilon\overline{\mathbf{w}}, \varepsilon\mathbf{a}), \qquad \varepsilon \in \{-1, 1\}. \tag{166}$$

(iv) If σ is neither even-linear nor constant-odd, corresponding nonconstant neurons have identical parameters.

If $\mathbf{r} \neq \mathbf{0}$, each parameterization also contains exactly one constant neuron. Its hidden bias b may be chosen arbitrarily subject to $\sigma(b) \neq 0$, with readout

$$\mathbf{a} = \frac{\mathbf{r}(\mathbf{0})}{\sigma(b)}. \tag{167}$$

If $\mathbf{r} = \mathbf{0}$, neither parameterization contains a constant neuron. Conversely, the transformations above, together with the stated freedom in the constant neuron, preserve both the symmetry orbit and irreducibility.

Proof. By Lemma D.7, both parameterizations contain exactly one neuron from each essential parameter class and, when $\mathbf{r} \neq \mathbf{0}$, exactly one constant neuron. We may therefore match their nonconstant neurons by parameter class and, when present, their constant neurons, which determines a permutation of neurons. Fix an essential class q, and let

$$(\overline{\mathbf{w}}^\star_{\boldsymbol{\theta}}, \mathbf{a}^\star_{\boldsymbol{\theta}}) \qquad \text{and} \qquad (\overline{\mathbf{w}}^\star_{\boldsymbol{\xi}}, \mathbf{a}^\star_{\boldsymbol{\xi}}) \tag{168}$$

denote the unique neurons of $\boldsymbol{\theta}^\star$ and $\boldsymbol{\xi}^\star$, respectively, belonging to q. Since both parameterizations lie in the same orbit, these neurons must realize the same aggregate $\boldsymbol{\beta}_q$. We now determine the resulting freedom in their parameters.

Suppose first that σ is purely even and positively homogeneous of degree 1. Since the two incoming-parameter vectors belong to the same class, there exists $t \in \mathbb{R}_{\neq 0}$ such that

$$\overline{\mathbf{w}}^\star_{\boldsymbol{\xi}} = t\overline{\mathbf{w}}^\star_{\boldsymbol{\theta}}. \tag{169}$$

Since a single neuron contributes $\|\overline{\mathbf{w}}\|\,\mathbf{a}$ to the class aggregate $\boldsymbol{\beta}_q$, equality of the aggregates gives

$$\|\overline{\mathbf{w}}^\star_{\boldsymbol{\theta}}\|\,\mathbf{a}^\star_{\boldsymbol{\theta}} = \|\overline{\mathbf{w}}^\star_{\boldsymbol{\xi}}\|\,\mathbf{a}^\star_{\boldsymbol{\xi}} = |t|\|\overline{\mathbf{w}}^\star_{\boldsymbol{\theta}}\|\,\mathbf{a}^\star_{\boldsymbol{\xi}}, \tag{170}$$

and hence

$$\mathbf{a}^\star_{\boldsymbol{\xi}} = |t|^{-1}\mathbf{a}^\star_{\boldsymbol{\theta}}. \tag{171}$$

If σ is purely even but not positively homogeneous of degree 1, the two elements of q differ only by sign, so

$$\overline{\mathbf{w}}^\star_{\boldsymbol{\xi}} = \varepsilon\overline{\mathbf{w}}^\star_{\boldsymbol{\theta}}, \qquad \varepsilon \in \{-1, 1\}. \tag{172}$$

The aggregate of a single neuron equals its readout, and therefore $\mathbf{a}^\star_{\boldsymbol{\xi}} = \mathbf{a}^\star_{\boldsymbol{\theta}}$.

If σ is purely odd, the incoming parameters in q again differ by a sign, so

$$\overline{\mathbf{w}}^\star_{\boldsymbol{\xi}} = \varepsilon\overline{\mathbf{w}}^\star_{\boldsymbol{\theta}}, \qquad \varepsilon \in \{-1, 1\}, \tag{173}$$

but reversing the incoming parameters of a neuron reverses the sign with which that neuron's readout enters the aggregate $\boldsymbol{\beta}_q$. Equality of aggregates therefore requires

$$\mathbf{a}^\star_{\boldsymbol{\xi}} = \varepsilon\mathbf{a}^\star_{\boldsymbol{\theta}}. \tag{174}$$

Finally, if σ is neither even-linear nor constant-odd, every parameter class is a singleton. Hence,

$$\overline{\mathbf{w}}^\star_{\boldsymbol{\xi}} = \overline{\mathbf{w}}^\star_{\boldsymbol{\theta}}, \qquad \mathbf{a}^\star_{\boldsymbol{\xi}} = \mathbf{a}^\star_{\boldsymbol{\theta}}. \tag{175}$$

When $\mathbf{r} \neq \mathbf{0}$, the unique constant neuron is solely responsible for the global residual and therefore satisfies

$$\mathbf{a}\sigma(b) = \mathbf{r}(\mathbf{0}) \neq \mathbf{0}. \tag{176}$$

Consequently, $\sigma(b) \neq 0$ and

$$\mathbf{a} = \frac{\mathbf{r}(\mathbf{0})}{\sigma(b)}, \tag{177}$$

while b may otherwise be chosen arbitrarily.

Conversely, each listed transformation preserves the corresponding class aggregate $\boldsymbol{\beta}_q$, and the stated freedom in the constant neuron preserves the global residual $\mathbf{r}$. By Proposition B.15, the transformed parameterization therefore remains in the same symmetry orbit. The transformations also preserve the structure characterized in Lemma D.7, and hence preserve irreducibility. □

The preceding corollary compares irreducible parameterizations within the same orbit. To prove Corollary 4.3 from the main text, we now compare an irreducible parameterization with an arbitrary parameterization in its orbit. The structural characterization in Lemma D.7 then allows us to show that every feature row of the former occurs in the latter up to nonzero scaling.

Proof of Corollary 4.3. By Lemma D.7, $\boldsymbol{\theta}^\star$ contains exactly one neuron from each essential parameter class and, when the global residual $\mathbf{r}$ is nonzero, exactly one constant neuron.

We first associate the neurons of $\boldsymbol{\theta}$ belonging to essential parameter classes with the corresponding neurons of $\boldsymbol{\theta}^\star$. Fix an essential class q, and let j denote the unique neuron of $\boldsymbol{\theta}^\star$ belonging to q. The feature row of every neuron i of $\boldsymbol{\theta}$ in the same class is a nonzero scalar multiple of $\mathbf{H}^\star_{j,:}$. Specifically, for singleton classes and purely even activations that are not positively 1-homogeneous, we set $\alpha_i = 1$. For purely odd activations, we set $\alpha_i = 1$ or $\alpha_i = -1$ according to whether the incoming parameters of neurons i and j agree or are opposite. Finally, for purely even positively 1-homogeneous activations, we set

$$\alpha_i \coloneqq \frac{\|\overline{\mathbf{w}}_i\|}{\|\overline{\mathbf{w}}^\star_j\|} > 0. \tag{178}$$

In each case,

$$\mathbf{H}_{i,:} = \alpha_i \mathbf{H}^\star_{j,:}. \tag{179}$$

Since every essential class is represented in $\boldsymbol{\theta}$, each nonconstant row of $\mathbf{H}^\star$ has at least one associated neuron.

Suppose next that $\boldsymbol{\theta}^\star$ contains a constant neuron j. Then, $\mathbf{r} \neq \mathbf{0}$ and $\sigma(b^\star_j) \neq 0$. Because, under the assumptions of Corollary 4.3, only constant neurons contribute to the global residual (see the proof of Lemma D.7), $\boldsymbol{\theta}$ must contain at least one constant neuron with nonzero output. For every such neuron i, associate it with neuron j and set

$$\alpha_i \coloneqq \frac{\sigma(b_i)}{\sigma(b^\star_j)} \neq 0. \tag{180}$$

Its feature row then satisfies

$$\mathbf{H}_{i,:} = \alpha_i \mathbf{H}^\star_{j,:}. \tag{181}$$

Consequently, every row of $\mathbf{H}^\star$ has at least one associated neuron. Let $\nu_j > 0$ denote the number of neurons associated with the jth row, for $j = 1, \dots, N^\star_h$.

It remains to account for the neurons of $\boldsymbol{\theta}$ not yet associated with a row of $\mathbf{H}^\star$. Let $\mathbf{U}$ contain each distinct feature row among these remaining neurons exactly once, and let K denote the number of such rows. Associate every remaining neuron i with its matching row of $\mathbf{U}$ and set $\alpha_i \coloneqq 1$. For $k = 1, \dots, K$, let $\nu_{N^\star_h + k} > 0$ denote the number of neurons associated with the kth row of $\mathbf{U}$.

Every neuron of $\boldsymbol{\theta}$ has now been associated exactly once, so

$$\boldsymbol{\nu} \in \mathbb{N}_{>0}^{N^\star_h + K}, \qquad \sum_{j=1}^{N^\star_h + K} \nu_j = N_h. \tag{182}$$

After ordering the neurons according to these associations, duplicating the rows of $[\mathbf{H}^{\star\top}, \mathbf{U}^\top]^\top$ according to $\boldsymbol{\nu}$ and applying the corresponding nonzero scaling factors yields Equation (11). By construction, the rows of $\mathbf{U}$ are pairwise distinct. □

E Representational geometry and similarity within symmetry orbits

This appendix provides proofs and extensions of the results in Section 5. We first derive the decomposition of the RSM into essential and auxiliary contributions and establish their independent reweighting under positive scaling (Appendix E.1). We then prove that the spread of attainable representational similarity scores with respect to a fixed reference increases strictly with width until it reaches its supremum across widths (Appendix E.2). Next, we show that the limiting lower and upper similarity bounds are independent of the symmetry orbit and are attained at finite width within every orbit (Appendix E.3). Finally, we characterize when similarity to a reference geometry can approach 1, establish conditions under which arbitrary reference geometries admit such alignment, and exhibit an obstruction showing that alignment need not be possible in general (Appendix E.4).

E.1 RSM decomposition and positive reweighting

We first extend the decomposition in Proposition 5.1 for positively 1-homogeneous activations to the remaining activation classes.

Proposition E.1 (RSM decomposition)**.** *Consider the factorization of* $\mathbf{H}$ *in Proposition D.6. Denote the rows of* $\mathbf{F}_{\mathcal{E},\mathcal{I}}$ *and* $\mathbf{U}$ *by* $\mathbf{f}_\ell^\top$ *and* $\mathbf{u}_k^\top$*, respectively. Then,*

$$\mathbf{H}^\top\mathbf{H} = \underbrace{\sum_{\ell=1}^{|\mathcal{I}|}\gamma_\ell \mathbf{f}_\ell\mathbf{f}_\ell^\top}_{\text{essential}} + \underbrace{\sum_{k=1}^{K}\gamma_{|\mathcal{I}|+k}\mathbf{u}_k\mathbf{u}_k^\top}_{\text{auxiliary}}, \qquad \boldsymbol{\gamma} \coloneqq \mathbf{D}_{\boldsymbol{\nu}}^\top\boldsymbol{\alpha}^2 \in \mathbb{R}_{>0}^{|\mathcal{I}|+K}, \tag{183}$$

where $\boldsymbol{\alpha}^2$ *denotes the Hadamard square. If* σ *is* not *positively* 1*-homogeneous, the factorization may be chosen with* $\boldsymbol{\alpha} = \mathbf{1}$*, in which case* $\boldsymbol{\gamma} = \boldsymbol{\nu}$*.*

Proof. By Lemma D.5,

$$\mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\alpha}^2)\mathbf{D}_{\boldsymbol{\nu}} = \operatorname{diag}(\mathbf{D}_{\boldsymbol{\nu}}^\top\boldsymbol{\alpha}^2) = \operatorname{diag}(\boldsymbol{\gamma}). \tag{184}$$

Hence,

$$\mathbf{H}^\top\mathbf{H} = \begin{bmatrix}\mathbf{F}_{\mathcal{E},\mathcal{I}}\\ \mathbf{U}\end{bmatrix}^\top \operatorname{diag}(\boldsymbol{\gamma}) \begin{bmatrix}\mathbf{F}_{\mathcal{E},\mathcal{I}}\\ \mathbf{U}\end{bmatrix} = \sum_{\ell=1}^{|\mathcal{I}|}\gamma_\ell\mathbf{f}_\ell\mathbf{f}_\ell^\top + \sum_{k=1}^{K}\gamma_{|\mathcal{I}|+k}\mathbf{u}_k\mathbf{u}_k^\top. \tag{185}$$

When $\boldsymbol{\alpha} = \mathbf{1}$, the identity $\mathbf{D}_{\boldsymbol{\nu}}^\top\mathbf{1} = \boldsymbol{\nu}$ gives the final claim. □

For the activation classes covered by Corollary 4.3, the same calculation yields a decomposition in terms of the features of any irreducible parameterization.

Corollary E.2 (RSM decomposition relative to an irreducible parameterization)**.** *Suppose that* σ *is neither even-linear nor constant-odd, or is purely even or purely odd. Let* $\mathbf{H}^\star$*,* $\mathbf{U}$*,* $\boldsymbol{\alpha}$*, and* $\boldsymbol{\nu}$ *be as in Corollary 4.3, and denote the rows of* $\mathbf{H}^\star$ *and* $\mathbf{U}$ *by* ${\mathbf{v}_j^\star}^\top$ *and* $\mathbf{u}_k^\top$*, respectively. Then,*

$$\mathbf{H}^\top\mathbf{H} = \underbrace{\sum_{j=1}^{N_h^\star}\gamma_j\mathbf{v}_j^\star{\mathbf{v}_j^\star}^\top}_{\text{essential}} + \underbrace{\sum_{k=1}^{K}\gamma_{N_h^\star+k}\mathbf{u}_k\mathbf{u}_k^\top}_{\text{auxiliary}}, \qquad \boldsymbol{\gamma} \coloneqq \mathbf{D}_{\boldsymbol{\nu}}^\top\boldsymbol{\alpha}^2 \in \mathbb{R}_{>0}^{N_h^\star+K}. \tag{186}$$

Proof. The stated decomposition follows directly by applying the Gram calculation from the proof of Proposition E.1 to the factorization of $\mathbf{H}$ in Equation (11). □

To prove Proposition 5.1, it remains to establish the independent reweighting claim for positively 1-homogeneous activations.

Proof of Proposition 5.1. Fix a parameterization $\boldsymbol{\theta}$ whose factorization of $\mathbf{H}^\top\mathbf{H}$ in Equation (12) has effective weights $\boldsymbol{\gamma}$, and let $\boldsymbol{\omega} \in \mathbb{R}_{>0}^{|\mathcal{I}|+K}$ be any target effective weights. We construct a symmetry-equivalent parameterization $\boldsymbol{\xi} \sim \boldsymbol{\theta}$ whose factorization of $\mathbf{H}^\top\mathbf{H}$ in Equation (12) has effective weights

$$\mathbf{D}_{\boldsymbol{\nu}}^\top\widetilde{\boldsymbol{\alpha}}^2 = \boldsymbol{\omega}, \qquad \widetilde{\boldsymbol{\alpha}} \in \mathbb{R}_{>0}^{N_h}, \tag{187}$$

while all other components of the factorization remain unchanged.

Let

$$t_\ell \coloneqq \sqrt{\frac{\omega_\ell}{\gamma_\ell}} > 0, \qquad \ell = 1,\dots,|\mathcal{I}|+K, \tag{188}$$

and apply the positive-scaling symmetry

$$(\mathbf{w}_j, b_j, \mathbf{a}_j) \mapsto (t_\ell\mathbf{w}_j, t_\ell b_j, t_\ell^{-1}\mathbf{a}_j) \tag{189}$$

to the jth neuron's parameters for every $j = 1,\dots,N_h$ such that $(\mathbf{D}_{\boldsymbol{\nu}})_{j\ell} = 1$. This yields a parameterization $\boldsymbol{\xi}$ that is symmetry-equivalent to $\boldsymbol{\theta}$. The scaling vector $\widetilde{\boldsymbol{\alpha}} \in \mathbb{R}_{>0}^{N_h}$ defined by

$$\widetilde{\alpha}_j \coloneqq t_\ell\alpha_j, \qquad j = 1,\dots,N_h, \tag{190}$$

where, for each j, ℓ is the unique index such that $(\mathbf{D}_{\boldsymbol{\nu}})_{j\ell} = 1$, satisfies

$$\left(\mathbf{D}_{\boldsymbol{\nu}}^\top \widetilde{\boldsymbol{\alpha}}^2\right)_\ell = t_\ell^2 \left(\mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2\right)_\ell = t_\ell^2 \gamma_\ell = \omega_\ell. \tag{191}$$

At the same time, the feature matrices $\mathbf{F}_{\mathcal{E},\mathcal{I}}$ and $\mathbf{U}$, the duplication pattern $\boldsymbol{\nu}$, and the width N_h are unchanged, as required. □

E.2 Strict growth until saturation

We adopt the notation $\mathcal{R}$, $\mathcal{R}_{N_h}$, $\mathcal{S}(\mathbf{N})$, $\mathcal{S}_{N_h}(\mathbf{N})$, $\Delta_{N_h}(\mathbf{N})$, and $\Delta_\infty(\mathbf{N})$ from Section 5.2. Throughout Appendices E.2 to E.4, we assume that σ is positively 1-homogeneous.

Pearson correlation as cosine similarity. Let $\mathcal{H}$ denote the space of symmetric matrices with zero diagonal and zero mean across off-diagonal entries:

$$\mathcal{H} := \{\mathbf{M} \in \mathbb{R}^{P\times P} \mid \mathbf{M} = \mathbf{M}^\top,\ \operatorname{diag}(\mathbf{M}) = \mathbf{0},\ \mathbf{1}^\top \mathbf{M} \mathbf{1} = 0\}. \tag{192}$$

The orthogonal projection $\Pi_{\mathcal{H}}$ of a symmetric matrix $\mathbf{M}$ onto $\mathcal{H}$ under the Frobenius inner product is given by

$$\Pi_{\mathcal{H}}(\mathbf{M}) = \mathbf{M} - \operatorname{diag}(\operatorname{diag}(\mathbf{M})) - \frac{\langle \mathbf{M}, \mathbf{1}\mathbf{1}^\top - \mathbf{I}\rangle_F}{P(P-1)}(\mathbf{1}\mathbf{1}^\top - \mathbf{I}). \tag{193}$$

Thus, $\Pi_{\mathcal{H}}$ sets the diagonal to zero and centers the off-diagonal entries by subtracting their mean. For any symmetric matrix $\mathbf{M}$ with nonconstant strict upper-triangular entries, define its normalized projection onto $\mathcal{H}$ by

$$\mathbf{M}^\circ := \frac{\Pi_{\mathcal{H}}(\mathbf{M})}{\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F}. \tag{194}$$

Under this normalization, the Pearson correlation $\rho(\mathbf{M}, \mathbf{N})$ between the strict upper-triangular entries of two symmetric matrices $\mathbf{M}$ and $\mathbf{N}$ coincides with the cosine similarity induced by the Frobenius inner product on $\mathcal{H}$.

Lemma E.3 (Pearson correlation as cosine similarity)**.** *Let $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{P\times P}$ be symmetric matrices with nonconstant strict upper-triangular entries. Then, their Pearson correlation is*

$$\rho(\mathbf{M}, \mathbf{N}) = \frac{\langle \Pi_{\mathcal{H}}(\mathbf{M}), \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F}{\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F \|\Pi_{\mathcal{H}}(\mathbf{N})\|_F} = \langle \mathbf{M}^\circ, \mathbf{N}^\circ\rangle_F. \tag{195}$$

Proof. Let

$$\hat{m} := \frac{2}{P(P-1)} \sum_{\mu<\nu} M_{\mu\nu}, \qquad \hat{n} := \frac{2}{P(P-1)} \sum_{\mu<\nu} N_{\mu\nu} \tag{196}$$

denote the means of the strict upper-triangular entries of $\mathbf{M}$ and $\mathbf{N}$, respectively. By definition of $\Pi_{\mathcal{H}}$,

$$\left[\Pi_{\mathcal{H}}(\mathbf{M})\right]_{\mu\nu} = M_{\mu\nu} - \hat{m}, \qquad \left[\Pi_{\mathcal{H}}(\mathbf{N})\right]_{\mu\nu} = N_{\mu\nu} - \hat{n}, \qquad \mu \neq \nu, \tag{197}$$

and

$$\left[\Pi_{\mathcal{H}}(\mathbf{M})\right]_{\mu\mu} = \left[\Pi_{\mathcal{H}}(\mathbf{N})\right]_{\mu\mu} = 0. \tag{198}$$

Hence, the Pearson correlation between the strict upper-triangular entries of $\mathbf{M}$ and $\mathbf{N}$ is

$$\rho(\mathbf{M}, \mathbf{N}) = \frac{\sum_{\mu<\nu}(M_{\mu\nu} - \hat{m})(N_{\mu\nu} - \hat{n})}{\left(\sum_{\mu<\nu}(M_{\mu\nu} - \hat{m})^2\right)^{1/2} \left(\sum_{\mu<\nu}(N_{\mu\nu} - \hat{n})^2\right)^{1/2}}. \tag{199}$$

Since the projected matrices are symmetric with zero diagonal,

$$\langle \Pi_{\mathcal{H}}(\mathbf{M}), \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F = 2 \sum_{\mu<\nu} (M_{\mu\nu} - \hat{m})(N_{\mu\nu} - \hat{n}), \tag{200}$$

and similarly

$$\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F^2 = 2 \sum_{\mu<\nu} (M_{\mu\nu} - \hat{m})^2, \qquad \|\Pi_{\mathcal{H}}(\mathbf{N})\|_F^2 = 2 \sum_{\mu<\nu} (N_{\mu\nu} - \hat{n})^2. \tag{201}$$

The factors of 2 therefore cancel in the cosine ratio, yielding

$$\rho(\mathbf{M}, \mathbf{N}) = \frac{\langle \Pi_{\mathcal{H}}(\mathbf{M}), \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F}{\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F \|\Pi_{\mathcal{H}}(\mathbf{N})\|_F} = \langle \mathbf{M}^\circ, \mathbf{N}^\circ\rangle_F, \tag{202}$$

where the final identity follows from Equation (194). □

Scaling and feature addition. The following construction shows that, for any parameterization $\boldsymbol{\theta}$ of width N_h, there exists a parameterization $\boldsymbol{\xi}$ of width $N_h + 1$ in the same symmetry orbit whose RSM equals a rescaled version of the RSM of $\boldsymbol{\theta}$ plus an independently weighted rank-one contribution from an added neuron with vanishing readout.

Lemma E.4 (RSM scaling and zero-readout feature addition)**.** *Let $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ with RSM $\mathbf{M}_{\boldsymbol{\theta}}$, and let $\mathbf{u}^\top = \sigma(\overline{\mathbf{w}}^\top \overline{\mathbf{X}})$ be any single-neuron feature row on the inputs $\mathbf{X}$. For every $\lambda > 0$ and $t \geq 0$, there exists $\boldsymbol{\xi} \in \mathcal{O}_{N_h+1}$ such that*

$$\mathbf{M}_{\boldsymbol{\xi}} = \lambda \mathbf{M}_{\boldsymbol{\theta}} + t\mathbf{u}\mathbf{u}^\top. \tag{203}$$

Moreover, $\lambda \mathbf{M}_{\boldsymbol{\theta}}$ is realizable within $\mathcal{O}_{N_h}$.

Proof. First, apply the positive-scaling symmetry to every neuron of $\boldsymbol{\theta}$:

$$(\overline{\mathbf{w}}_j, \mathbf{a}_j) \mapsto \big(\sqrt{\lambda}\,\overline{\mathbf{w}}_j, \lambda^{-1/2}\mathbf{a}_j\big), \qquad j = 1, \ldots, N_h. \tag{204}$$

This preserves the realized function while multiplying every feature by $\sqrt{\lambda}$. Hence, the hidden-activation matrix is multiplied by $\sqrt{\lambda}$ and its RSM by λ. In particular, $\lambda \mathbf{M}_{\boldsymbol{\theta}}$ is realizable within $\mathcal{O}_{N_h}$.

Next, append a single neuron with parameters

$$(\sqrt{t}\,\overline{\mathbf{w}}, \mathbf{0}), \qquad t \geq 0. \tag{205}$$

If $t > 0$, positive 1-homogeneity gives the feature $\sqrt{t}\,\mathbf{u}^\top$. If $t = 0$, the feature vanishes because positive 1-homogeneity forces $\sigma(0) = 0$. The appended neuron forms a singleton zero-neuron group when its incoming weights are nonzero and is a zero-readout constant neuron otherwise. In either case, its addition preserves the symmetry orbit, while its contribution to the RSM is $t\mathbf{u}\mathbf{u}^\top$. Therefore,

$$\mathbf{M}_{\boldsymbol{\xi}} = \lambda \mathbf{M}_{\boldsymbol{\theta}} + t\mathbf{u}\mathbf{u}^\top, \tag{206}$$

as claimed. □

The preceding lemma provides the mechanism by which increasing width enlarges the set of realizable geometries: increasing the width by one can leave an existing RSM unchanged or augment it by an independently weighted rank-one contribution from any realizable single-neuron feature. We use this to show that the sets of realizable RSMs and attainable similarity scores are nested with increasing width, and that the spread $\Delta_{N_h}(\mathbf{N})$ of attainable similarity scores increases strictly whenever it has not yet reached its orbit-wide limit $\Delta_\infty(\mathbf{N})$.

Proof of Proposition 5.2. We divide the proof into three steps.

Nesting. Let $\mathbf{M} \in \mathcal{R}_{N_h}$. Setting $\lambda = 1$ and $t = 0$ in Lemma E.4 yields a parameterization in $\mathcal{O}_{N_h+1}$ with the same RSM $\mathbf{M}$. Hence, every RSM realizable at width N_h is also realizable at width $N_h + 1$, so

$$\mathcal{R}_{N_h} \subseteq \mathcal{R}_{N_h+1}. \tag{207}$$

Applying $\rho(\,\cdot\,, \mathbf{N})$ to these sets then gives

$$\mathcal{S}_{N_h}(\mathbf{N}) \subseteq \mathcal{S}_{N_h+1}(\mathbf{N}). \tag{208}$$

Strict improvement of the upper bound. Fix a width N_h, and suppose

$$s \coloneqq \sup \mathcal{S}_{N_h}(\mathbf{N}) < \sup \mathcal{S}(\mathbf{N}). \tag{209}$$

Then, there exists an RSM $\mathbf{B}$ realized by a parameterization in the same symmetry orbit at some finite width L such that

$$\rho(\mathbf{B}, \mathbf{N}) > s. \tag{210}$$

Necessarily, $L > N_h$, since nesting would otherwise imply $\rho(\mathbf{B}, \mathbf{N}) \in \mathcal{S}_{N_h}(\mathbf{N})$. Let $\mathbf{v}_1^\top, \ldots, \mathbf{v}_L^\top$ denote the feature rows of the parameterization giving rise to $\mathbf{B}$, that is, $\mathbf{B} = \sum_j \mathbf{v}_j \mathbf{v}_j^\top$. By linearity of the projection $\Pi_{\mathcal{H}}$, we have

$$\mathbf{Y} \coloneqq \Pi_{\mathcal{H}}(\mathbf{B}) = \sum_{j=1}^{L} \mathbf{G}_j, \qquad \mathbf{G}_j \coloneqq \Pi_{\mathcal{H}}(\mathbf{v}_j \mathbf{v}_j^\top). \tag{211}$$

Since $\rho(\mathbf{B}, \mathbf{N}) = \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F / \|\mathbf{Y}\|_F$ by Lemma E.3, the choice of $\mathbf{B}$ implies

$$\langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F > s \|\mathbf{Y}\|_F. \tag{212}$$

We will show that a single feature from this wider parameterization can already be added to a suitable width-N_h parameterization to obtain similarity strictly greater than s at width $N_h + 1$.

Case $s \geq 0$. Choose a sequence $\mathbf{M}_i \in \mathcal{R}_{N_h}$ such that

$$\lim_{i \to \infty} \rho(\mathbf{M}_i, \mathbf{N}) = s. \tag{213}$$

By the same-width scaling construction in Lemma E.4, each $\mathbf{M}_i$ may be rescaled within $\mathcal{R}_{N_h}$, without changing its similarity to $\mathbf{N}$, so that $\|\Pi_{\mathcal{H}}(\mathbf{M}_i)\|_F = 1$, and hence $\Pi_{\mathcal{H}}(\mathbf{M}_i) = \mathbf{M}_i^\circ$. Consequently, the sequence $\mathbf{M}_i^\circ$ lies on the unit sphere of the finite-dimensional vector space $\mathcal{H}$ defined in Equation (192). By compactness of this sphere, there exist a subsequence $\mathbf{M}_{i_k}^\circ$ and a matrix $\mathbf{Q} \in \mathcal{H}$ such that $\lim_{k \to \infty} \mathbf{M}_{i_k}^\circ = \mathbf{Q}$. By continuity of the Frobenius norm,

$$\|\mathbf{Q}\|_F = \lim_{k \to \infty} \|\mathbf{M}_{i_k}^\circ\|_F = 1. \tag{214}$$

Similarly, continuity of the Frobenius inner product gives

$$\langle \mathbf{Q}, \mathbf{N}^\circ \rangle_F = \lim_{k \to \infty} \langle \mathbf{M}_{i_k}^\circ, \mathbf{N}^\circ \rangle_F = \lim_{k \to \infty} \rho(\mathbf{M}_{i_k}, \mathbf{N}) = s, \tag{215}$$

where the penultimate equality follows from Lemma E.3. Since $s \geq 0$ and $\|\mathbf{Q}\|_F = 1$, the Cauchy–Schwarz inequality gives

$$\langle \mathbf{Y}, \mathbf{N}^\circ - s\mathbf{Q} \rangle_F = \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F - s \langle \mathbf{Y}, \mathbf{Q} \rangle_F \geq \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F - s \|\mathbf{Y}\|_F > 0. \tag{216}$$

Using $\mathbf{Y} = \sum_j \mathbf{G}_j$, we therefore have

$$\sum_{j=1}^{L} \langle \mathbf{G}_j, \mathbf{N}^\circ - s\mathbf{Q} \rangle_F > 0. \tag{217}$$

Hence, there exists at least one index j such that

$$\langle \mathbf{G}_j, \mathbf{N}^\circ - s\mathbf{Q} \rangle_F > 0. \tag{218}$$

For this index j, define

$$\phi(t) := \frac{\langle \mathbf{Q} + t\mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|\mathbf{Q} + t\mathbf{G}_j\|_F}. \tag{219}$$

Since $\|\mathbf{Q}\|_F = 1$ and $\langle \mathbf{Q}, \mathbf{N}^\circ \rangle_F = s$, we have $\phi(0) = s$. Differentiating at $t = 0$ gives

$$\phi'(0) = \langle \mathbf{G}_j, \mathbf{N}^\circ - s\mathbf{Q} \rangle_F > 0. \tag{220}$$

Hence, we can choose $t > 0$ sufficiently small that

$$\frac{\langle \mathbf{Q} + t\mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|\mathbf{Q} + t\mathbf{G}_j\|_F} > s. \tag{221}$$

Continuity and $\mathbf{M}_{i_k}^\circ \to \mathbf{Q}$ imply that, for all sufficiently large k,

$$\frac{\langle \mathbf{M}_{i_k}^\circ + t\mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|\mathbf{M}_{i_k}^\circ + t\mathbf{G}_j\|_F} > s. \tag{222}$$

By linearity of $\Pi_{\mathcal{H}}$ and the definition of $\mathbf{G}_j$,

$$\Pi_{\mathcal{H}}(\mathbf{M}_{i_k} + t\mathbf{v}_j\mathbf{v}_j^\top) = \mathbf{M}_{i_k}^\circ + t\mathbf{G}_j. \tag{223}$$

Therefore, Lemma E.3 gives, for all sufficiently large k,

$$\rho(\mathbf{M}_{i_k} + t\mathbf{v}_j\mathbf{v}_j^\top, \mathbf{N}) = \frac{\langle \mathbf{M}_{i_k}^\circ + t\mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|\mathbf{M}_{i_k}^\circ + t\mathbf{G}_j\|_F} > s. \tag{224}$$

By Lemma E.4, each such $\mathbf{M}_{i_k} + t\mathbf{v}_j\mathbf{v}_j^\top$ is realizable within $\mathcal{O}_{N_h+1}$. Hence,

$$\sup \mathcal{S}_{N_h+1}(\mathbf{N}) > s = \sup \mathcal{S}_{N_h}(\mathbf{N}). \tag{225}$$

Case $s < 0$. We first show that there exists an index j with $\mathbf{G}_j \neq \mathbf{0}$ such that

$$\frac{\langle \mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|\mathbf{G}_j\|_F} > s. \tag{226}$$

Suppose, to the contrary, that every nonzero $\mathbf{G}_j$ satisfies $\langle \mathbf{G}_j, \mathbf{N}^\circ \rangle_F \leq s\|\mathbf{G}_j\|_F$. The same inequality holds trivially when $\mathbf{G}_j = \mathbf{0}$. Summing over j and using $\mathbf{Y} = \sum_j \mathbf{G}_j$ gives

$$\langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F \leq s \sum_{j=1}^{L} \|\mathbf{G}_j\|_F \leq s\|\mathbf{Y}\|_F, \tag{227}$$

where the second inequality follows from the triangle inequality and $s < 0$. This contradicts Equation (212), establishing that there does exist an index j with $\mathbf{G}_j \neq \mathbf{0}$ satisfying Equation (226). Fix any $\mathbf{M}_0 \in \mathcal{R}_{N_h}$, and choose an index j satisfying Equation (226). For all sufficiently large $t > 0$, linearity of $\Pi_{\mathcal{H}}$ and Lemma E.3 give

$$\rho(\mathbf{M}_0 + t\mathbf{v}_j\mathbf{v}_j^\top, \mathbf{N}) = \frac{\langle \Pi_{\mathcal{H}}(\mathbf{M}_0) + t\mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|\Pi_{\mathcal{H}}(\mathbf{M}_0) + t\mathbf{G}_j\|_F} = \frac{\langle t^{-1}\Pi_{\mathcal{H}}(\mathbf{M}_0) + \mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|t^{-1}\Pi_{\mathcal{H}}(\mathbf{M}_0) + \mathbf{G}_j\|_F}. \tag{228}$$

Clearly, $\lim_{t\to\infty} t^{-1}\Pi_{\mathcal{H}}(\mathbf{M}_0) = \mathbf{0}$. Since $\mathbf{G}_j \neq \mathbf{0}$, continuity of the Frobenius inner product and norm therefore gives

$$\lim_{t\to\infty} \rho(\mathbf{M}_0 + t\mathbf{v}_j\mathbf{v}_j^\top, \mathbf{N}) = \frac{\langle \mathbf{G}_j, \mathbf{N}^\circ \rangle_F}{\|\mathbf{G}_j\|_F} > s, \tag{229}$$

where the final inequality follows from Equation (226). Hence, for some sufficiently large finite $t > 0$, we have $\rho(\mathbf{M}_0 + t\mathbf{v}_j\mathbf{v}_j^\top, \mathbf{N}) > s$. By Lemma E.4, this RSM is realizable within $\mathcal{O}_{N_h+1}$. Therefore,

$$\sup \mathcal{S}_{N_h+1}(\mathbf{N}) > s = \sup \mathcal{S}_{N_h}(\mathbf{N}). \tag{230}$$

In both cases, we have shown that $\sup \mathcal{S}_{N_h}(\mathbf{N}) < \sup \mathcal{S}(\mathbf{N})$ implies $\sup \mathcal{S}_{N_h}(\mathbf{N}) < \sup \mathcal{S}_{N_h+1}(\mathbf{N})$. Conversely, if $\sup \mathcal{S}_{N_h}(\mathbf{N}) = \sup \mathcal{S}(\mathbf{N})$, then

$$\mathcal{S}_{N_h}(\mathbf{N}) \subseteq \mathcal{S}_{N_h+1}(\mathbf{N}) \subseteq \mathcal{S}(\mathbf{N}) \tag{231}$$

implies

$$\sup \mathcal{S}_{N_h+1}(\mathbf{N}) = \sup \mathcal{S}_{N_h}(\mathbf{N}) = \sup \mathcal{S}(\mathbf{N}). \tag{232}$$

Therefore,

$$\sup \mathcal{S}_{N_h+1}(\mathbf{N}) > \sup \mathcal{S}_{N_h}(\mathbf{N}) \quad \Longleftrightarrow \quad \sup \mathcal{S}_{N_h}(\mathbf{N}) < \sup \mathcal{S}(\mathbf{N}). \tag{233}$$

Lower bound and spread. The argument establishing Equation (233) depends on the reference geometry $\mathbf{N}$ only through $\mathbf{N}^\circ$. Replacing $\mathbf{N}^\circ$ by $-\mathbf{N}^\circ$ negates every similarity score, since

$$\langle \mathbf{M}^\circ, -\mathbf{N}^\circ \rangle_F = -\rho(\mathbf{M}, \mathbf{N}). \tag{234}$$

Repeating the argument leading to Equation (233) with $-\mathbf{N}^\circ$ in place of $\mathbf{N}^\circ$ therefore gives

$$\inf \mathcal{S}_{N_h+1}(\mathbf{N}) < \inf \mathcal{S}_{N_h}(\mathbf{N}) \quad \Longleftrightarrow \quad \inf \mathcal{S}_{N_h}(\mathbf{N}) > \inf \mathcal{S}(\mathbf{N}). \tag{235}$$

Finally, by definition,

$$\Delta_\infty(\mathbf{N}) - \Delta_{N_h}(\mathbf{N}) = (\sup \mathcal{S}(\mathbf{N}) - \sup \mathcal{S}_{N_h}(\mathbf{N})) + (\inf \mathcal{S}_{N_h}(\mathbf{N}) - \inf \mathcal{S}(\mathbf{N})). \tag{236}$$

Both terms are nonnegative because, by definition, $\mathcal{S}_{N_h}(\mathbf{N}) \subseteq \mathcal{S}(\mathbf{N})$. Hence, $\Delta_{N_h}(\mathbf{N}) < \Delta_\infty(\mathbf{N})$ if and only if at least one of the upper or lower similarity bounds has not reached its orbit-wide limit. Similarly,

$$\begin{aligned}\Delta_{N_h+1}(\mathbf{N}) - \Delta_{N_h}(\mathbf{N}) = (&\sup \mathcal{S}_{N_h+1}(\mathbf{N}) - \sup \mathcal{S}_{N_h}(\mathbf{N})) \\ &+ (\inf \mathcal{S}_{N_h}(\mathbf{N}) - \inf \mathcal{S}_{N_h+1}(\mathbf{N})).\end{aligned} \tag{237}$$

Again, both terms are nonnegative, since $\mathcal{S}_{N_h}(\mathbf{N}) \subseteq \mathcal{S}_{N_h+1}(\mathbf{N})$ by Equation (208). By Equations (233) and (235), at least one term is strictly positive if and only if at least one of the corresponding orbit-wide bounds has not yet been reached. Therefore,

$$\Delta_{N_h+1}(\mathbf{N}) > \Delta_{N_h}(\mathbf{N}) \quad \Longleftrightarrow \quad \Delta_{N_h}(\mathbf{N}) < \Delta_\infty(\mathbf{N}), \tag{238}$$

completing the proof. □

E.3 Function-independent similarity limits

Proposition 5.3 establishes that the limiting lower and upper similarity bounds are independent of the symmetry orbit and that the corresponding fixed-width bounds stabilize at finite width. We prove this result by characterizing the closure of the set of normalized projections onto $\mathcal{H}$ of RSMs realizable within a symmetry orbit. We write cl for closure with respect to the Frobenius norm and $\overline{\mathrm{cone}}(\mathcal{A})$ for the closed conic hull of a set $\mathcal{A}$.

The closed projected feature cone. Define the set of projected single-neuron contributions

$$\mathcal{G}_{\sigma,\mathbf{X}} \coloneqq \{\Pi_{\mathcal{H}}(\mathbf{u}\mathbf{u}^\top) \mid \mathbf{u}^\top = \sigma(\overline{\mathbf{w}}^\top \overline{\mathbf{X}}),\ \overline{\mathbf{w}} \in \mathbb{R}^{N_i+1}\}. \tag{239}$$

The closed projected feature cone is then

$$\mathcal{C}_{\sigma,\mathbf{X}} \coloneqq \overline{\mathrm{cone}}\big(\mathcal{G}_{\sigma,\mathbf{X}}\big) \subseteq \mathcal{H}. \tag{240}$$

By definition, both $\mathcal{G}_{\sigma,\mathbf{X}}$ and $\mathcal{C}_{\sigma,\mathbf{X}}$ depend only on σ and $\mathbf{X}$, and not on any particular symmetry orbit. Let

$$d \coloneqq \dim \operatorname{span} \mathcal{C}_{\sigma,\mathbf{X}} \leq \dim \mathcal{H} = \frac{P(P-1)}{2} - 1. \tag{241}$$

Since $\mathcal{S}(\mathbf{N})$ is nonempty, there exists $\mathbf{M} = \mathbf{H}^\top \mathbf{H} \in \mathcal{R}$ with nonconstant strict upper-triangular entries, and hence $\Pi_{\mathcal{H}}(\mathbf{M}) \neq \mathbf{0}$. Writing the rows of $\mathbf{H}$ as $\mathbf{u}_j^\top$ gives

$$\Pi_{\mathcal{H}}(\mathbf{M}) = \sum_j \Pi_{\mathcal{H}}(\mathbf{u}_j \mathbf{u}_j^\top) \in \mathrm{cone}(\mathcal{G}_{\sigma,\mathbf{X}}) \subseteq \mathcal{C}_{\sigma,\mathbf{X}}. \tag{242}$$

Thus, $\mathcal{C}_{\sigma,\mathbf{X}} \neq \{\mathbf{0}\}$, so $d \geq 1$.

The following lemma characterizes the closure of the set of normalized projections onto $\mathcal{H}$ of RSMs attainable within any symmetry orbit. In particular, for every $N_h \geq N_h^\star + d$, every unit-norm element of $\mathcal{C}_{\sigma,\mathbf{X}}$ lies in the closure of the normalized projections onto $\mathcal{H}$ of RSMs realizable at width N_h.

Lemma E.5 (Closure of normalized projected geometries within a symmetry orbit)**.** *Fix a symmetry orbit $\mathcal{O}$, and let $N_h^\star$ denote the minimum width among parameterizations in $\mathcal{O}$. Then,*

$$\mathrm{cl}\{\mathbf{M}^\circ \mid \mathbf{M} \in \mathcal{R}_{N_h}\} = \{\mathbf{Y} \in \mathcal{C}_{\sigma,\mathbf{X}} \mid \|\mathbf{Y}\|_F = 1\}, \qquad N_h \geq N_h^\star + d. \tag{243}$$

Proof. We prove the set equality in two steps, showing that each set is contained in the other.

Cone membership. Fix N_h and let $\mathbf{M} \in \mathcal{R}_{N_h}$. Let $\mathbf{v}_1^\top, \ldots, \mathbf{v}_{N_h}^\top$ denote the feature rows of a parameterization realizing $\mathbf{M}$, i.e., $\mathbf{M} = \sum_j \mathbf{v}_j \mathbf{v}_j^\top$. Linearity of $\Pi_{\mathcal{H}}$ gives

$$\Pi_{\mathcal{H}}(\mathbf{M}) = \sum_{j=1}^{N_h} \Pi_{\mathcal{H}}(\mathbf{v}_j \mathbf{v}_j^\top) \in \mathrm{cone}(\mathcal{G}_{\sigma,\mathbf{X}}) \subseteq \mathcal{C}_{\sigma,\mathbf{X}}. \tag{244}$$

Because $\mathbf{M} \in \mathcal{R}_{N_h}$, its strict upper-triangular entries are nonconstant, so $\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F > 0$. Since $\mathcal{C}_{\sigma,\mathbf{X}}$ is a cone,

$$\mathbf{M}^\circ = \frac{\Pi_{\mathcal{H}}(\mathbf{M})}{\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F} \in \mathcal{C}_{\sigma,\mathbf{X}}, \qquad \|\mathbf{M}^\circ\|_F = 1. \tag{245}$$

The set

$$\{\mathbf{Y} \in \mathcal{C}_{\sigma,\mathbf{X}} \mid \|\mathbf{Y}\|_F = 1\} \tag{246}$$

is closed because it is the intersection of $\mathcal{C}_{\sigma,\mathbf{X}}$ and the unit sphere in $\mathcal{H}$, both of which are closed. Therefore,

$$\mathrm{cl}\{\mathbf{M}^\circ \mid \mathbf{M} \in \mathcal{R}_{N_h}\} \subseteq \{\mathbf{Y} \in \mathcal{C}_{\sigma,\mathbf{X}} \mid \|\mathbf{Y}\|_F = 1\}. \tag{247}$$

Approximation by at most d feature contributions. Fix $\mathbf{Y} \in \mathcal{C}_{\sigma,\mathbf{X}}$ with $\|\mathbf{Y}\|_F = 1$. Since $\mathcal{C}_{\sigma,\mathbf{X}} = \overline{\mathrm{cone}}(\mathcal{G}_{\sigma,\mathbf{X}})$, there exists a sequence

$$\{\mathbf{Y}_i\}_{i=1}^{\infty} \subseteq \mathrm{cone}(\mathcal{G}_{\sigma,\mathbf{X}}) \qquad \text{such that} \qquad \lim_{i\to\infty} \mathbf{Y}_i = \mathbf{Y}. \tag{248}$$

Moreover, $\operatorname{span} \mathcal{G}_{\sigma,\mathbf{X}} = \operatorname{span} \mathcal{C}_{\sigma,\mathbf{X}}$, because the finite-dimensional space $\operatorname{span} \mathcal{G}_{\sigma,\mathbf{X}}$ is closed and therefore contains the closure of $\mathrm{cone}(\mathcal{G}_{\sigma,\mathbf{X}})$. Hence, the generators lie in a d-dimensional vector space. By the conic version of Carathéodory's theorem, each $\mathbf{Y}_i$ can therefore be expressed using at most d generators:

$$\mathbf{Y}_i = \sum_{j=1}^{d} t_{ij}\mathbf{G}_{ij}, \qquad t_{ij} \geq 0, \tag{249}$$

where $\mathbf{G}_{ij} \in \mathcal{G}_{\sigma,\mathbf{X}}$. If fewer than d generators are required, we set the remaining coefficients to zero. Since each $\mathbf{G}_{ij} \in \mathcal{G}_{\sigma,\mathbf{X}}$, for every i and j there exists $\overline{\mathbf{w}}_{ij} \in \mathbb{R}^{N_i+1}$ such that

$$\mathbf{G}_{ij} = \Pi_{\mathcal{H}}(\mathbf{u}_{ij}\mathbf{u}_{ij}^{\top}), \qquad \mathbf{u}_{ij}^{\top} = \sigma(\overline{\mathbf{w}}_{ij}^{\top}\overline{\mathbf{X}}). \tag{250}$$

Realization within the orbit. Choose an irreducible parameterization $\boldsymbol{\theta}^{\star} \in \mathcal{O}_{N_h^{\star}}$ with RSM $\mathbf{M}_{\boldsymbol{\theta}^{\star}}$, and let $\{\lambda_i\}_{i=1}^{\infty}$ be any positive sequence with $\lambda_i \to 0$. Iterating Lemma E.4 yields, for every $i \geq 1$, a parameterization in $\mathcal{O}_{N_h^{\star}+d}$ with RSM

$$\mathbf{M}_i \coloneqq \lambda_i \mathbf{M}_{\boldsymbol{\theta}^{\star}} + \sum_{j=1}^{d} t_{ij}\mathbf{u}_{ij}\mathbf{u}_{ij}^{\top}. \tag{251}$$

By linearity of $\Pi_{\mathcal{H}}$ and the definition of $\mathbf{Y}_i$,

$$\Pi_{\mathcal{H}}(\mathbf{M}_i) = \lambda_i \Pi_{\mathcal{H}}(\mathbf{M}_{\boldsymbol{\theta}^{\star}}) + \sum_{j=1}^{d} t_{ij}\Pi_{\mathcal{H}}(\mathbf{u}_{ij}\mathbf{u}_{ij}^{\top}) = \lambda_i \Pi_{\mathcal{H}}(\mathbf{M}_{\boldsymbol{\theta}^{\star}}) + \mathbf{Y}_i. \tag{252}$$

Since $\lambda_i \to 0$ and $\mathbf{Y}_i \to \mathbf{Y}$, we have $\Pi_{\mathcal{H}}(\mathbf{M}_i) \to \mathbf{Y}$. By continuity of the Frobenius norm,

$$\lim_{i\to\infty} \|\Pi_{\mathcal{H}}(\mathbf{M}_i)\|_F = \|\mathbf{Y}\|_F = 1. \tag{253}$$

Hence, $\Pi_{\mathcal{H}}(\mathbf{M}_i) \neq \mathbf{0}$ for all sufficiently large i, so $\mathbf{M}_i \in \mathcal{R}_{N_h^{\star}+d}$ for all such i. Moreover,

$$\lim_{i\to\infty} \mathbf{M}_i^{\circ} = \lim_{i\to\infty} \frac{\Pi_{\mathcal{H}}(\mathbf{M}_i)}{\|\Pi_{\mathcal{H}}(\mathbf{M}_i)\|_F} = \mathbf{Y}, \tag{254}$$

again by continuity. Therefore,

$$\mathbf{Y} \in \mathrm{cl}\{\mathbf{M}^{\circ} \mid \mathbf{M} \in \mathcal{R}_{N_h^{\star}+d}\}. \tag{255}$$

This proves the reverse inclusion at width $N_h^{\star} + d \leq N_h$. By nesting, $\mathcal{R}_{N_h^{\star}+d} \subseteq \mathcal{R}_{N_h}$ for every $N_h \geq N_h^{\star} + d$, so the same inclusion holds at every such width. Together with the set inclusion in Equation (247), this proves Equation (243). □

The characterization in Lemma E.5 now yields both claims of Proposition 5.3.

Proof of Proposition 5.3. Fix a symmetry orbit $\mathcal{O}$, and let $N_h^{\star}$ denote the minimum width among parameterizations in $\mathcal{O}$. Let

$$\mathcal{K}_{\sigma,\mathbf{X}} \coloneqq \{\mathbf{Y} \in \mathcal{C}_{\sigma,\mathbf{X}} \mid \|\mathbf{Y}\|_F = 1\}. \tag{256}$$

The set $\mathcal{K}_{\sigma,\mathbf{X}}$ is the intersection of the closed set $\mathcal{C}_{\sigma,\mathbf{X}}$ with the unit sphere in the finite-dimensional space $\mathcal{H}$, and is therefore compact. Since $\mathcal{C}_{\sigma,\mathbf{X}}$ is a nontrivial cone, $\mathcal{K}_{\sigma,\mathbf{X}}$ is also nonempty.

Fix $N_h \geq N_h^{\star} + d$. By Lemma E.5,

$$\mathrm{cl}\{\mathbf{M}^{\circ} \mid \mathbf{M} \in \mathcal{R}_{N_h}\} = \mathcal{K}_{\sigma,\mathbf{X}}. \tag{257}$$

By Lemma E.3, $\rho(\mathbf{M},\mathbf{N}) = \langle \mathbf{M}^{\circ}, \mathbf{N}^{\circ} \rangle_F$. Since $\mathcal{K}_{\sigma,\mathbf{X}}$ is compact and nonempty, the continuous map $\mathbf{Y} \mapsto \langle \mathbf{Y}, \mathbf{N}^{\circ} \rangle_F$ attains its minimum and maximum on $\mathcal{K}_{\sigma,\mathbf{X}}$. By density of the realizable normalized projections $\mathbf{M}^{\circ}$ in $\mathcal{K}_{\sigma,\mathbf{X}}$,

$$\inf \mathcal{S}_{N_h}(\mathbf{N}) = \inf_{\mathbf{M}\in\mathcal{R}_{N_h}} \rho(\mathbf{M},\mathbf{N}) = \inf_{\mathbf{M}\in\mathcal{R}_{N_h}} \langle \mathbf{M}^{\circ}, \mathbf{N}^{\circ} \rangle_F = \min_{\mathbf{Y}\in\mathcal{K}_{\sigma,\mathbf{X}}} \langle \mathbf{Y}, \mathbf{N}^{\circ} \rangle_F, \tag{258}$$

and similarly,

$$\sup \mathcal{S}_{N_h}(\mathbf{N}) = \max_{\mathbf{Y} \in \mathcal{K}_{\sigma,\mathbf{X}}} \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F. \tag{259}$$

By the *cone membership* part of the proof of Lemma E.5, every $\mathbf{M} \in \mathcal{R}$ satisfies $\mathbf{M}^\circ \in \mathcal{K}_{\sigma,\mathbf{X}}$. Hence, $\{\mathbf{M}^\circ \mid \mathbf{M} \in \mathcal{R}\} \subseteq \mathcal{K}_{\sigma,\mathbf{X}}$, and thus

$$\min_{\mathbf{Y} \in \mathcal{K}_{\sigma,\mathbf{X}}} \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F \leq \inf \mathcal{S}(\mathbf{N}) \leq \inf \mathcal{S}_{N_h}(\mathbf{N}) = \min_{\mathbf{Y} \in \mathcal{K}_{\sigma,\mathbf{X}}} \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F, \tag{260}$$

where the second inequality follows from $\mathcal{S}_{N_h}(\mathbf{N}) \subseteq \mathcal{S}(\mathbf{N})$, forcing

$$\inf \mathcal{S}(\mathbf{N}) = \inf \mathcal{S}_{N_h}(\mathbf{N}) = \min_{\mathbf{Y} \in \mathcal{K}_{\sigma,\mathbf{X}}} \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F. \tag{261}$$

The same argument yields

$$\sup \mathcal{S}(\mathbf{N}) = \sup \mathcal{S}_{N_h}(\mathbf{N}) = \max_{\mathbf{Y} \in \mathcal{K}_{\sigma,\mathbf{X}}} \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F. \tag{262}$$

The right-hand sides in Equations (261) and (262) depend only on σ, $\mathbf{X}$, and $\mathbf{N}$, and are independent of the symmetry orbit $\mathcal{O}$. Hence, $\inf \mathcal{S}(\mathbf{N})$ and $\sup \mathcal{S}(\mathbf{N})$ are identical across all symmetry orbits, and the identities

$$\inf \mathcal{S}_{N_h}(\mathbf{N}) = \inf \mathcal{S}(\mathbf{N}), \qquad \sup \mathcal{S}_{N_h}(\mathbf{N}) = \sup \mathcal{S}(\mathbf{N}), \qquad N_h \geq N_h^\star + d \tag{263}$$

yield

$$\Delta_{N_h}(\mathbf{N}) = \Delta_\infty(\mathbf{N}), \qquad N_h \geq N_h^\star + d. \tag{264}$$

Hence, every symmetry orbit reaches the common limiting similarity bounds $\inf \mathcal{S}(\mathbf{N})$ and $\sup \mathcal{S}(\mathbf{N})$, and therefore the common limiting spread $\Delta_\infty(\mathbf{N})$, at finite width. □

E.4 Alignment with reference geometries

The preceding characterization of the closure of the set of normalized projected RSMs immediately yields a necessary and sufficient condition under which RSMs realizable within a symmetry orbit can approach perfect representational similarity with a fixed reference geometry $\mathbf{N}$ at sufficient width.

Corollary E.6 (Alignment with reference geometries)**.** *Fix a symmetry orbit $\mathcal{O}$ with activation σ, and let $N_h^\star$ denote the minimum width among parameterizations in $\mathcal{O}$. Then,*

$$\sup \mathcal{S}(\mathbf{N}) = 1 \quad \Longleftrightarrow \quad \mathbf{N}^\circ \in \mathcal{C}_{\sigma,\mathbf{X}}. \tag{265}$$

Moreover, for every $N_h \geq N_h^\star + d$,

$$\sup \mathcal{S}_{N_h}(\mathbf{N}) = 1 \quad \Longleftrightarrow \quad \mathbf{N}^\circ \in \mathcal{C}_{\sigma,\mathbf{X}}. \tag{266}$$

Proof. Recall from the proof of Proposition 5.3 that $\mathcal{K}_{\sigma,\mathbf{X}}$ denotes the set of unit-norm elements of $\mathcal{C}_{\sigma,\mathbf{X}}$. By Equation (262),

$$\sup \mathcal{S}(\mathbf{N}) = \sup \mathcal{S}_{N_h}(\mathbf{N}) = \max_{\mathbf{Y} \in \mathcal{K}_{\sigma,\mathbf{X}}} \langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F, \qquad N_h \geq N_h^\star + d. \tag{267}$$

Since $\|\mathbf{Y}\|_F = 1$ for $\mathbf{Y} \in \mathcal{K}_{\sigma,\mathbf{X}}$, and $\|\mathbf{N}^\circ\|_F = 1$, the Cauchy–Schwarz inequality gives $\langle \mathbf{Y}, \mathbf{N}^\circ \rangle_F \leq 1$, with equality if and only if $\mathbf{Y} = \mathbf{N}^\circ$. Hence, the maximum equals 1 if and only if $\mathbf{N}^\circ \in \mathcal{C}_{\sigma,\mathbf{X}}$, proving both claims. □

Corollary E.6 is agnostic to the source of the reference RSM, which may be derived from neural data, another neural network, or any other computational model. The result depends *only* on the normalized projection $\mathbf{N}^\circ$ and the cone $\mathcal{C}_{\sigma,\mathbf{X}}$ determined by σ and $\mathbf{X}$. Intuitively, this cone consists of the projected RSMs induced by features realizable by neurons with activation function σ on the inputs $\mathbf{X}$.

References generated with the same activation. Suppose the reference RSM is generated by a network that uses the same activation function σ and is evaluated on the same inputs $\mathbf{X}$. Its normalized projection $\mathbf{N}^\circ$ then belongs to $\mathcal{C}_{\sigma,\mathbf{X}}$, since the neuron-wise decomposition of $\mathbf{N}^\circ$ expresses it as a conic combination of projected single-neuron contributions. If the reference network has width L, the projection admits such a representation with at most $\min\{d, L\}$ contributions.

Proposition E.7 (Same-activation reference geometries)**.** *Suppose that* $\mathbf{N}$ *is the RSM of a width-*L *network with activation* σ*, evaluated on* $\mathbf{X}$*. Then, for every symmetry orbit* $\mathcal{O}$ *with activation* σ*,*

$$\sup \mathcal{S}(\mathbf{N}) = 1. \tag{268}$$

Moreover, if $N_h^\star$ *denotes the minimum width among parameterizations in* $\mathcal{O}$*, then*

$$\sup \mathcal{S}_{N_h}(\mathbf{N}) = 1, \qquad N_h \geq N_h^\star + \min\{d, L\}. \tag{269}$$

Proof. Let $\mathbf{v}_j$ denote the feature of the jth hidden neuron of the reference network on the inputs $\mathbf{X}$. Then, $\mathbf{N} = \sum_{j=1}^{L} \mathbf{v}_j \mathbf{v}_j^\top$. By linearity of the projection $\Pi_{\mathcal{H}}$,

$$\Pi_{\mathcal{H}}(\mathbf{N}) = \sum_{j=1}^{L} \Pi_{\mathcal{H}}(\mathbf{v}_j \mathbf{v}_j^\top) \in \operatorname{cone}(\mathcal{G}_{\sigma,\mathbf{X}}) \subseteq \mathcal{C}_{\sigma,\mathbf{X}}. \tag{270}$$

By assumption, the strict upper-triangular entries of $\mathbf{N}$ are nonconstant, so $\Pi_{\mathcal{H}}(\mathbf{N}) \neq \mathbf{0}$. Because $\mathcal{C}_{\sigma,\mathbf{X}}$ is a cone,

$$\mathbf{N}^\circ = \frac{\Pi_{\mathcal{H}}(\mathbf{N})}{\|\Pi_{\mathcal{H}}(\mathbf{N})\|_F} \in \mathcal{C}_{\sigma,\mathbf{X}}. \tag{271}$$

Hence, Corollary E.6 establishes Equation (268) and yields

$$\sup \mathcal{S}_{N_h}(\mathbf{N}) = 1, \qquad N_h \geq N_h^\star + d. \tag{272}$$

To establish Equation (269), it remains to show that $\sup \mathcal{S}_{N_h^\star + L}(\mathbf{N}) = 1$. Choose an irreducible parameterization $\boldsymbol{\theta}^\star$ of width $N_h^\star$ in $\mathcal{O}$, and denote its RSM by $\mathbf{M}_0$. Iteratively applying Lemma E.4 yields a symmetry-equivalent parameterization $\boldsymbol{\xi} \sim \boldsymbol{\theta}^\star$ of width $N_h^\star + L$ with RSM

$$\mathbf{M}_\lambda = \lambda \mathbf{M}_0 + \sum_{j=1}^{L} \mathbf{v}_j \mathbf{v}_j^\top = \lambda \mathbf{M}_0 + \mathbf{N}, \qquad \lambda > 0. \tag{273}$$

By linearity of $\Pi_{\mathcal{H}}$,

$$\lim_{\lambda \downarrow 0} \Pi_{\mathcal{H}}(\mathbf{M}_\lambda) = \lim_{\lambda \downarrow 0} \big(\lambda \Pi_{\mathcal{H}}(\mathbf{M}_0) + \Pi_{\mathcal{H}}(\mathbf{N})\big) = \Pi_{\mathcal{H}}(\mathbf{N}). \tag{274}$$

Since $\Pi_{\mathcal{H}}(\mathbf{N}) \neq \mathbf{0}$, Equation (274) implies that $\Pi_{\mathcal{H}}(\mathbf{M}_\lambda) \neq \mathbf{0}$ for all sufficiently small $\lambda > 0$, and hence $\mathbf{M}_\lambda \in \mathcal{R}_{N_h^\star + L}$. The normalized projections therefore satisfy

$$\lim_{\lambda \downarrow 0} \mathbf{M}_\lambda^\circ = \frac{\Pi_{\mathcal{H}}(\mathbf{N})}{\|\Pi_{\mathcal{H}}(\mathbf{N})\|_F} = \mathbf{N}^\circ. \tag{275}$$

By Lemma E.3 and continuity of the Frobenius inner product, we obtain

$$\lim_{\lambda \downarrow 0} \rho(\mathbf{M}_\lambda, \mathbf{N}) = \lim_{\lambda \downarrow 0} \langle \mathbf{M}_\lambda^\circ, \mathbf{N}^\circ \rangle_F = \langle \mathbf{N}^\circ, \mathbf{N}^\circ \rangle_F = 1. \tag{276}$$

Since Pearson correlation is bounded above by 1, this implies that $\sup \mathcal{S}_{N_h^\star + L}(\mathbf{N}) = 1$. Combining this with Equation (272) gives $\sup \mathcal{S}_{N_h^\star + \min\{d,L\}}(\mathbf{N}) = 1$. The nesting of the attainable similarity sets $\mathcal{S}_{N_h}(\mathbf{N})$ as width increases (Proposition 5.2) then establishes Equation (269), completing the proof. □

Proposition E.7 shows once more that representational geometry can vary substantially while the computed function remains unchanged. Reference features can be added with zero readouts and amplified to make the RSM arbitrarily well aligned with $\mathbf{N}$, regardless of the function computed by the reference network. Figure 2 illustrates this mechanism by adding, with zero readout, a feature from an opposite-family solution (i.e., from the reference network giving rise to $\mathbf{N}$) in Figure 1. Duplication and scaling then amplify the feature's rank-one RSM contribution, which is nearly perfectly correlated with $\mathbf{N}$, until it dominates the full RSM without changing the realized function.

A sufficient condition for arbitrary-reference alignment. The preceding result shows that, when the reference geometry $\mathbf{N}$ is generated by a network with activation σ, similarity to that reference can approach 1 within every symmetry orbit with the same activation. For an *arbitrary* reference geometry, however, membership of $\mathbf{N}^\circ$ in $\mathcal{C}_{\sigma,\mathbf{X}}$ is not automatic. We next give a sufficient condition under which $\mathcal{C}_{\sigma,\mathbf{X}}$ does contain the normalized projection of every reference geometry.

Proposition E.8 (Affinely independent probe inputs)**.** *Suppose that* $\operatorname{rank}(\overline{\mathbf{X}}) = P$*. Then,* $\mathcal{C}_{\sigma,\mathbf{X}} = \mathcal{H}$*. Consequently, for any symmetry orbit* $\mathcal{O}$ *with activation* σ *and minimum width* $N_h^\star$*,*

$$\inf \mathcal{S}(\mathbf{N}) = -1, \qquad \sup \mathcal{S}(\mathbf{N}) = 1. \tag{277}$$

Moreover, for every $N_h \geq N_h^\star + \frac{P(P-1)}{2} - 1$*,*

$$\inf \mathcal{S}_{N_h}(\mathbf{N}) = -1, \qquad \sup \mathcal{S}_{N_h}(\mathbf{N}) = 1. \tag{278}$$

Proof. Choose $z_0 \in \mathbb{R}$ such that $\sigma(z_0) \neq 0$. The rank assumption $\operatorname{rank}(\overline{\mathbf{X}}) = P$ implies that the map

$$\mathbb{R}^{N_i+1} \to \mathbb{R}^P, \quad \overline{\mathbf{w}} \mapsto \overline{\mathbf{X}}^\top \overline{\mathbf{w}} \tag{279}$$

is surjective. Hence, for any $\mathbf{v} \in \mathbb{R}^P_{\geq 0}$, there exists $\overline{\mathbf{w}} \in \mathbb{R}^{N_i+1}$ such that

$$\frac{z_0}{|\sigma(z_0)|}\mathbf{v}^\top = (\overline{\mathbf{X}}^\top \overline{\mathbf{w}})^\top = \overline{\mathbf{w}}^\top \overline{\mathbf{X}}. \tag{280}$$

Since $\mathbf{v}$ has nonnegative entries, positive 1-homogeneity gives

$$\sigma(\overline{\mathbf{w}}^\top \overline{\mathbf{X}}) = \frac{\sigma(z_0)}{|\sigma(z_0)|}\mathbf{v}^\top = \varepsilon \mathbf{v}^\top, \qquad \varepsilon \in \{-1, 1\}. \tag{281}$$

Hence,

$$\Pi_{\mathcal{H}}(\mathbf{v}\mathbf{v}^\top) \in \mathcal{G}_{\sigma,\mathbf{X}} \subseteq \mathcal{C}_{\sigma,\mathbf{X}}, \qquad \mathbf{v} \in \mathbb{R}^P_{\geq 0}. \tag{282}$$

Now, let $\mathbf{Y} \in \mathcal{H}$ be any symmetric matrix with zero diagonal and zero-mean off-diagonal entries. Set $c \coloneqq -\min_{\mu<\nu} Y_{\mu\nu}$, and let $\mathbf{e}^\mu \in \mathbb{R}^P$ denote the μth standard basis vector, $\mu = 1, \ldots, P$. The matrix

$$\mathbf{B} \coloneqq \sum_{\mu<\nu} (Y_{\mu\nu} + c)(\mathbf{e}^\mu + \mathbf{e}^\nu)(\mathbf{e}^\mu + \mathbf{e}^\nu)^\top \in \mathbb{R}^{P\times P} \tag{283}$$

is symmetric by construction and has off-diagonal entries

$$B_{\mu\nu} = Y_{\mu\nu} + c \geq 0, \qquad \mu \neq \nu. \tag{284}$$

Thus,

$$\frac{2}{P(P-1)} \sum_{\mu<\nu} B_{\mu\nu} = \frac{2}{P(P-1)} \sum_{\mu<\nu} (Y_{\mu\nu} + c) = c + \frac{2}{P(P-1)} \sum_{\mu<\nu} Y_{\mu\nu} = c, \tag{285}$$

since $\mathbf{Y}$ has zero-mean off-diagonal entries. Projecting $\mathbf{B}$ onto $\mathcal{H}$ sets its diagonal to zero and subtracts c from every off-diagonal entry. Therefore,

$$\Pi_{\mathcal{H}}(\mathbf{B}) = \mathbf{Y}. \tag{286}$$

As each vector $\mathbf{e}^\mu + \mathbf{e}^\nu \in \mathbb{R}^P_{\geq 0}$ has nonnegative entries, Equation (282) gives

$$\mathbf{Y} = \Pi_{\mathcal{H}}(\mathbf{B}) = \sum_{\mu<\nu} (Y_{\mu\nu} + c)\Pi_{\mathcal{H}}\big((\mathbf{e}^\mu + \mathbf{e}^\nu)(\mathbf{e}^\mu + \mathbf{e}^\nu)^\top\big) \in \mathcal{C}_{\sigma,\mathbf{X}}. \tag{287}$$

Hence, $\mathcal{H} \subseteq \mathcal{C}_{\sigma,\mathbf{X}}$, while the reverse inclusion follows directly from the definition of $\mathcal{C}_{\sigma,\mathbf{X}}$ in Equation (240).

It remains to establish Equation (278), since Equation (277) then follows from $\mathcal{S}_{N_h}(\mathbf{N}) \subseteq \mathcal{S}(\mathbf{N})$ and the fact that Pearson correlation takes values in $[-1, 1]$. For each $\varepsilon \in \{-1, 1\}$,

$$\varepsilon \mathbf{N}^\circ \in \mathcal{H} = \mathcal{C}_{\sigma,\mathbf{X}}, \qquad \|\varepsilon \mathbf{N}^\circ\|_F = 1, \tag{288}$$

and hence, by Lemma E.5,

$$\varepsilon \mathbf{N}^\circ \in \operatorname{cl}\{\mathbf{M}^\circ \mid \mathbf{M} \in \mathcal{R}_{N_h}\}, \qquad N_h \geq N_h^\star + d. \tag{289}$$

Thus, for every $\varepsilon \in \{-1, 1\}$ and every $N_h \geq N_h^\star + d$, there exists a sequence

$$\{\mathbf{M}_{i,\varepsilon}\}_{i=1}^{\infty} \subseteq \mathcal{R}_{N_h} \qquad \text{such that} \qquad \lim_{i\to\infty} \mathbf{M}_{i,\varepsilon}^{\circ} = \varepsilon \mathbf{N}^{\circ}. \tag{290}$$

By Lemma E.3 and continuity of the Frobenius inner product,

$$\lim_{i\to\infty} \rho(\mathbf{M}_{i,\varepsilon}, \mathbf{N}) = \lim_{i\to\infty} \langle \mathbf{M}_{i,\varepsilon}^{\circ}, \mathbf{N}^{\circ} \rangle_F = \varepsilon \langle \mathbf{N}^{\circ}, \mathbf{N}^{\circ} \rangle_F = \varepsilon. \tag{291}$$

Since Pearson correlation takes values in $[-1, 1]$, it follows that

$$\inf \mathcal{S}_{N_h}(\mathbf{N}) = -1, \qquad \sup \mathcal{S}_{N_h}(\mathbf{N}) = 1, \tag{292}$$

for every $N_h \geq N_h^\star + d$. Finally, $\mathcal{C}_{\sigma,\mathbf{X}} = \mathcal{H}$ implies

$$d = \dim \operatorname{span} \mathcal{C}_{\sigma,\mathbf{X}} = \dim \mathcal{H} = \frac{P(P-1)}{2} - 1, \tag{293}$$

which establishes Equation (278) and completes the proof. □

Proposition E.8 gives a sufficient condition under which every reference geometry admits similarities arbitrarily close to both -1 and 1. When this condition is not satisfied, however, the activation σ and probe inputs $\mathbf{X}$ can restrict the closed projected feature cone $\mathcal{C}_{\sigma,\mathbf{X}}$ and thereby prevent arbitrary alignment with particular reference geometries $\mathbf{N}$.

Example E.9 (A reference geometry with nonpositive similarity for ReLU networks)**.** Consider $\mathbf{N} = \mathbf{u}\mathbf{u}^\top$ with $\mathbf{u}^\top = (1, 0, 1)$ and probe inputs

$$\mathbf{X} = (-1, 0, 1) \in \mathbb{R}^3. \tag{294}$$

Clearly, $\operatorname{rank}(\overline{\mathbf{X}}) = 2 < 3$, so Proposition E.8 does not apply and $\sup \mathcal{S}(\mathbf{N}) = 1$ is not guaranteed. In fact, for every symmetry orbit $\mathcal{O}$ with ReLU activation, we will show that $\sup \mathcal{S}(\mathbf{N}) \leq 0$.

Write $\psi(z) = \max(0, z)$. On the inputs $\mathbf{X}$, a ReLU neuron with weight w and hidden bias b computes the feature $\mathbf{v} \in \mathbb{R}^3$ with entries

$$v_1 = \psi(b - w), \qquad v_2 = \psi(b), \qquad v_3 = \psi(b + w). \tag{295}$$

We first establish the inequality

$$v_1 v_2 - 2 v_1 v_3 + v_2 v_3 \geq 0. \tag{296}$$

If $v_1 v_3 = 0$, the inequality follows immediately from $v_1, v_2, v_3 \geq 0$. Otherwise, $v_1 > 0$ and $v_3 > 0$, so

$$b - w > 0, \qquad b + w > 0, \tag{297}$$

which implies $b > 0$. Hence, all three pre-activations are positive, and therefore

$$v_1 = b - w, \qquad v_2 = b, \qquad v_3 = b + w. \tag{298}$$

Thus,

$$v_1 v_2 - 2 v_1 v_3 + v_2 v_3 = (b - w)b - 2(b - w)(b + w) + b(b + w) = 2w^2 \geq 0, \tag{299}$$

proving Equation (296).

Now consider an arbitrary ReLU network of width N_h, and let $\mathbf{v}_j^\top$ denote the feature computed by neuron j on the probe inputs $\mathbf{X}$. The corresponding hidden-activation matrix $\mathbf{H}$ has rows $\mathbf{v}_j^\top$, so the resulting RSM satisfies

$$\mathbf{M} = \mathbf{H}^\top \mathbf{H} = \sum_{j=1}^{N_h} \mathbf{v}_j \mathbf{v}_j^\top. \tag{300}$$

Consequently,

$$M_{\mu\nu} = \sum_{j=1}^{N_h} v_{j\mu} v_{j\nu}, \qquad 1 \leq \mu < \nu \leq 3, \tag{301}$$

where $v_{j\mu}$ denotes the μth entry of $\mathbf{v}_j \in \mathbb{R}^3$. Hence, Equation (296) implies

$$M_{12} - 2M_{13} + M_{23} = \sum_{j=1}^{N_h} \big(v_{j1} v_{j2} - 2 v_{j1} v_{j3} + v_{j2} v_{j3} \big) \geq 0. \tag{302}$$

We have

$$\Pi_{\mathcal{H}}(\mathbf{N}) = \Pi_{\mathcal{H}}\left(\begin{bmatrix}1 & 0 & 1\\ 0 & 0 & 0\\ 1 & 0 & 1\end{bmatrix}\right) = \frac{1}{3}\begin{bmatrix}0 & -1 & 2\\ -1 & 0 & -1\\ 2 & -1 & 0\end{bmatrix}. \tag{303}$$

Since $\Pi_{\mathcal{H}}$ is the Frobenius-orthogonal projection onto $\mathcal{H}$ and $\Pi_{\mathcal{H}}(\mathbf{N}) \in \mathcal{H}$,

$$\langle \Pi_{\mathcal{H}}(\mathbf{M}), \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F = \langle \mathbf{M}, \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F = \frac{2}{3}(-M_{12} + 2M_{13} - M_{23}) \le 0 \tag{304}$$

by Equation (302). Whenever $\rho(\mathbf{M}, \mathbf{N})$ is defined, Lemma E.3 therefore yields

$$\rho(\mathbf{M}, \mathbf{N}) = \frac{\langle \Pi_{\mathcal{H}}(\mathbf{M}), \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F}{\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F \|\Pi_{\mathcal{H}}(\mathbf{N})\|_F} \le 0. \tag{305}$$

Since the ReLU network was arbitrary, $\sup \mathcal{S}(\mathbf{N}) \le 0$ for every symmetry orbit $\mathcal{O}$ with ReLU activation, as claimed.

The fact that no ReLU network can align positively with the reference geometry $\mathbf{N}$ from Example E.9 has a simple geometric explanation. The probe inputs satisfy

$$\mathbf{x}^2 = \frac{\mathbf{x}^1 + \mathbf{x}^3}{2}. \tag{306}$$

Hence, the pre-activation of any neuron on $\mathbf{x}^2$ is the average of its pre-activations on $\mathbf{x}^1$ and $\mathbf{x}^3$. If a neuron is active on both $\mathbf{x}^1$ and $\mathbf{x}^3$, then it is necessarily active on $\mathbf{x}^2$ as well, and linearity of ReLU on the positive half-line gives

$$v_2 = \frac{v_1 + v_3}{2}. \tag{307}$$

This relation is precisely what gives rise to Equation (296). If $v_1 v_3 = 0$, the inequality follows from nonnegativity of the ReLU features. Otherwise,

$$v_1 v_2 - 2v_1 v_3 + v_2 v_3 = \frac{(v_1 - v_3)^2}{2} \ge 0. \tag{308}$$

Thus, the affine dependence among the probe inputs imposes a linear constraint on the neuron-wise contributions to the pairwise similarities between two inputs. Summing this constraint over neurons yields

$$M_{13} \le \frac{M_{12} + M_{23}}{2} \tag{309}$$

for every RSM realizable by a ReLU network on $\mathbf{X}$. By contrast, the reference geometry $\mathbf{N}$ satisfies

$$N_{12} = N_{23} = 0, \qquad N_{13} = 1, \tag{310}$$

and therefore violates Equation (309). Moreover, the projected reference geometry $\Pi_{\mathcal{H}}(\mathbf{N})$ is a negative scalar multiple of the Frobenius normal

$$\mathbf{L} = \begin{bmatrix}0 & 1 & -2\\ 1 & 0 & 1\\ -2 & 1 & 0\end{bmatrix} \tag{311}$$

to the half-space defined by Equation (309), with $\Pi_{\mathcal{H}}(\mathbf{N}) = -1/3\,\mathbf{L}$. For every realizable RSM,

$$\langle \Pi_{\mathcal{H}}(\mathbf{M}), \mathbf{L}\rangle_F = 2(M_{12} - 2M_{13} + M_{23}) \ge 0, \tag{312}$$

and therefore

$$\langle \Pi_{\mathcal{H}}(\mathbf{M}), \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F \le 0. \tag{313}$$

The affine-independence assumption in Proposition E.8 rules out precisely the input dependencies that give rise to restrictions such as Equation (309). By preventing these dependencies from constraining the features, it allows the projected feature cone $\mathcal{C}_{\sigma,\mathbf{X}}$ to span all of $\mathcal{H}$ and thereby guarantees arbitrary alignment with any reference geometry.

F Identifiability through minimum-norm selection

This appendix provides proofs and extensions of the identifiability results in Section 6. We first characterize the representational geometries selected by minimum-weight-norm selection for positively 1-homogeneous activations, prove a sufficient condition for representational identifiability, and examine its implications for the analytical ReLU solutions illustrated in Figure 1 and discussed in detail in Appendix G (Appendix F.1). We then develop the corresponding minimum-representation-norm analysis, making explicit how the finite input set affects both representational identifiability and existence of a minimizer (Appendix F.2).

F.1 Minimum weight norm for positively 1-homogeneous activations

Throughout this subsection, fix a symmetry orbit $\mathcal{O}$ with positively 1-homogeneous activation $\sigma(z) = \delta|z| + mz$, essential parameter classes $\mathcal{E}$, unit-norm representatives $\overline{\mathbf{w}}_q$, orbit invariants $\boldsymbol{\beta}_q$, and residual $\mathbf{r}$. We adopt the feature vectors $\mathbf{f}_q^+$ and $\mathbf{f}_q^-$, $q \in \mathcal{E}$, and the sets $\mathcal{T}$ and $\mathcal{T}_{N_h}$ introduced in Section 6.

To establish Proposition 6.1 and Corollary 6.2, we first derive an orbit-wide lower bound on the weight-norm objective Ω_W and identify exactly when this bound is attained.

Lemma F.1 (Weight-norm lower bound and equality conditions)**.** *Every parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ satisfies*

$$\Omega_W(\boldsymbol{\theta}) \geq 2 \sum_{q \in \mathcal{E}} \|\boldsymbol{\beta}_q\|. \tag{314}$$

Equality holds if and only if the following three conditions are satisfied:

(i) Every neuron outside the essential parameter classes has parameters

$$\overline{\mathbf{w}}_j = \mathbf{0}, \qquad \mathbf{a}_j = \mathbf{0}. \tag{315}$$

(ii) Every neuron in an essential parameter class has equal incoming-parameter and readout norms:

$$\|\overline{\mathbf{w}}_j\| = \|\mathbf{a}_j\|, \qquad j \in \mathcal{J}_q, \quad q \in \mathcal{E}. \tag{316}$$

(iii) For every $q \in \mathcal{E}$, there exist coefficients $c_j \geq 0$, $j \in \mathcal{J}_q$, such that

$$\|\overline{\mathbf{w}}_j\| \mathbf{a}_j = c_j \boldsymbol{\beta}_q, \qquad \sum_{j \in \mathcal{J}_q} c_j = 1. \tag{317}$$

Proof. By definition of the weight-norm objective $\Omega_W(\boldsymbol{\theta})$ in Equation (18), we have

$$\Omega_W(\boldsymbol{\theta}) = \sum_{j=1}^{N_h} \big(\|\overline{\mathbf{w}}_j\|^2 + \|\mathbf{a}_j\|^2\big) \geq \sum_{q \in \mathcal{E}} \sum_{j \in \mathcal{J}_q} \big(\|\overline{\mathbf{w}}_j\|^2 + \|\mathbf{a}_j\|^2\big). \tag{318}$$

The AM–GM inequality gives

$$\sum_{q \in \mathcal{E}} \sum_{j \in \mathcal{J}_q} \big(\|\overline{\mathbf{w}}_j\|^2 + \|\mathbf{a}_j\|^2\big) \geq 2 \sum_{q \in \mathcal{E}} \sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\| \|\mathbf{a}_j\|. \tag{319}$$

For each essential parameter class $q \in \mathcal{E}$, absolute homogeneity of the Euclidean norm and the triangle inequality yield

$$\sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\| \|\mathbf{a}_j\| = \sum_{j \in \mathcal{J}_q} \big\|\|\overline{\mathbf{w}}_j\| \mathbf{a}_j\big\| \geq \Big\| \sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\| \mathbf{a}_j \Big\| = \|\boldsymbol{\beta}_q\|, \tag{320}$$

where the final equality follows from the definition of the class aggregate $\boldsymbol{\beta}_q$ in Equation (65). Together, Equations (318) to (320) establish the inequality in Equation (314).

Equality in Equation (314) holds if and only if equality holds in Equations (318) and (319) and, for every $q \in \mathcal{E}$, in Equation (320). It therefore remains to determine when each of these inequalities is tight.

Equality in Equation (318) holds if and only if every neuron outside the essential parameter classes contributes nothing to Ω_W. Since each such contribution is

$$\|\overline{\mathbf{w}}_j\|^2 + \|\mathbf{a}_j\|^2 \geq 0, \tag{321}$$

this is equivalent to

$$\overline{\mathbf{w}}_j = \mathbf{0}, \qquad \mathbf{a}_j = \mathbf{0}, \tag{322}$$

giving the first condition.

Equality in Equation (319) holds if and only if the AM–GM inequality is tight for every neuron in every essential parameter class. For each $j \in \mathcal{J}_q$,

$$\left(\|\overline{\mathbf{w}}_j\|^2 + \|\mathbf{a}_j\|^2\right) - 2\|\overline{\mathbf{w}}_j\|\|\mathbf{a}_j\| = \left(\|\overline{\mathbf{w}}_j\| - \|\mathbf{a}_j\|\right)^2, \tag{323}$$

so equality holds if and only if

$$\|\overline{\mathbf{w}}_j\| = \|\mathbf{a}_j\|, \qquad j \in \mathcal{J}_q, \quad q \in \mathcal{E}, \tag{324}$$

giving the second condition.

Finally, equality in Equation (320) is determined by equality in the triangle inequality. Since $\boldsymbol{\beta}_q \neq \mathbf{0}$ for every essential parameter class $q \in \mathcal{E}$, equality holds if and only if the vectors

$$\|\overline{\mathbf{w}}_j\|\mathbf{a}_j, \qquad j \in \mathcal{J}_q, \tag{325}$$

are nonnegative multiples of $\boldsymbol{\beta}_q$. Equivalently, there exist coefficients $c_j \geq 0$ such that

$$\|\overline{\mathbf{w}}_j\|\mathbf{a}_j = c_j\boldsymbol{\beta}_q, \qquad j \in \mathcal{J}_q. \tag{326}$$

Summing over $j \in \mathcal{J}_q$ and using the definition of the class aggregate gives

$$\boldsymbol{\beta}_q = \boldsymbol{\beta}_q \sum_{j\in\mathcal{J}_q} c_j. \tag{327}$$

Since $\boldsymbol{\beta}_q \neq \mathbf{0}$, we have $\sum_{j\in\mathcal{J}_q} c_j = 1$, giving the third condition.

Conversely, if the three stated conditions hold, then equality holds in Equations (318) to (320), and hence in Equation (314). □

The third condition in the preceding lemma also determines how the class aggregate $\boldsymbol{\beta}_q$ can be distributed between the two orientations $\mathcal{J}_q^+$ and $\mathcal{J}_q^-$ of an essential parameter class $q \in \mathcal{E}$.

Corollary F.2 (Readout allocation at the weight-norm lower bound)**.** *Suppose that $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ attains the lower bound in Equation (314). Then, for every essential parameter class $q \in \mathcal{E}$, there exists $t_q \in [0,1]$ such that*

$$\sum_{j\in\mathcal{J}_q^+} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = t_q\boldsymbol{\beta}_q, \qquad \sum_{j\in\mathcal{J}_q^-} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = (1-t_q)\boldsymbol{\beta}_q. \tag{328}$$

Proof. By Lemma F.1, there exist coefficients $c_j \geq 0$, $j \in \mathcal{J}_q$, such that

$$\|\overline{\mathbf{w}}_j\|\mathbf{a}_j = c_j\boldsymbol{\beta}_q, \qquad \sum_{j\in\mathcal{J}_q} c_j = 1. \tag{329}$$

Setting $t_q := \sum_{j\in\mathcal{J}_q^+} c_j$ gives $t_q \in [0,1]$ and $\sum_{j\in\mathcal{J}_q^-} c_j = 1 - t_q$, completing the proof. □

With the preceding lemma and corollary in place, it remains to establish two facts to prove Proposition 6.1. We first show that any parameterization attaining the lower bound in Equation (314) induces a split $\mathbf{t} \in \mathcal{T}_{N_h}$ and has an RSM of the form given in Equation (23). We then show that every split $\mathbf{t} \in \mathcal{T}_{N_h}$ is realized by a parameterization in $\mathcal{O}_{N_h}$ that attains this lower bound.

Before doing so, we make precise the width requirement in the definition of $\mathcal{T}_{N_h}$ in Section 6. A split $0 < t_q < 1$ requires at least two neurons from class q, one of each orientation, whereas a split $t_q \in \{0,1\}$ requires only one. Consequently, the minimum number of neurons from essential parameter classes needed to realize $\mathbf{t} \in \mathcal{T}$ is

$$\kappa(\mathbf{t}) := |\mathcal{E}| + |\{q \in \mathcal{E} \mid 0 < t_q < 1\}|. \tag{330}$$

With this notation, the set $\mathcal{T}_{N_h}$ of splits feasible at width N_h can be formally defined as

$$\mathcal{T}_{N_h} := \{\mathbf{t} \in \mathcal{T} \mid \kappa(\mathbf{t}) \leq N_h\}. \tag{331}$$

Proof of Proposition 6.1. We first show that every parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ attaining the lower bound in Equation (314) induces a split $\mathbf{t} \in \mathcal{T}_{N_h}$ for which its RSM is given by Equation (23).

Fix $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ attaining the lower bound in Equation (314). By Corollary F.2, for every $q \in \mathcal{E}$, there exists $t_q \in [0,1]$ such that the class aggregate $\boldsymbol{\beta}_q$ is allocated between the two orientations $\mathcal{J}_q^+$ and $\mathcal{J}_q^-$ according to t_q and $1-t_q$. By the first condition in Lemma F.1, every neuron outside the essential parameter classes has parameters $\overline{\mathbf{w}}_j = \mathbf{0}$ and $\mathbf{a}_j = \mathbf{0}$. Hence, the global residual $\mathbf{r}$ in Equation (66) is determined entirely by essential parameter classes, with each class $q \in \mathcal{E}$ contributing $m(2t_q - 1)\boldsymbol{\beta}_q \overline{\mathbf{w}}_q^\top \overline{\mathbf{x}}$ by Equation (21). Therefore,

$$\mathbf{r}(\mathbf{x}) = m \sum_{q \in \mathcal{E}} (2t_q - 1)\boldsymbol{\beta}_q \overline{\mathbf{w}}_q^\top \overline{\mathbf{x}}. \tag{332}$$

Since $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ has residual $\mathbf{r}$, the definition of $\mathcal{T}$ in Equation (22) therefore gives $\mathbf{t} \in \mathcal{T}$. Since $\mathbf{t}$ is induced by a parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ of width N_h, we have $\mathbf{t} \in \mathcal{T}_{N_h}$.

We next determine the RSM induced by $\boldsymbol{\theta}$. By the second and third conditions in Lemma F.1, for every $j \in \mathcal{J}_q$, there exists $c_j \geq 0$ such that

$$\|\overline{\mathbf{w}}_j\|^2 = \|\overline{\mathbf{w}}_j\| \|\mathbf{a}_j\| = \big| \|\overline{\mathbf{w}}_j\| \mathbf{a}_j \big| = c_j \|\boldsymbol{\beta}_q\|. \tag{333}$$

Let $\varepsilon \in \{+,-\}$. Since $\overline{\mathbf{w}}_q$ has unit norm, positive 1-homogeneity implies that neuron $j \in \mathcal{J}_q^\varepsilon$ has feature row

$$\sigma(\overline{\mathbf{w}}_j^\top \overline{\mathbf{X}}) = \|\overline{\mathbf{w}}_j\| \, \sigma(\varepsilon \overline{\mathbf{w}}_q^\top \overline{\mathbf{X}}) = \|\overline{\mathbf{w}}_j\| (\mathbf{f}_q^\varepsilon)^\top. \tag{334}$$

Its contribution to $\mathbf{H}^\top \mathbf{H}$ is therefore

$$\|\overline{\mathbf{w}}_j\| (\mathbf{f}_q^\varepsilon) \|\overline{\mathbf{w}}_j\| (\mathbf{f}_q^\varepsilon)^\top = c_j \|\boldsymbol{\beta}_q\| \mathbf{f}_q^\varepsilon (\mathbf{f}_q^\varepsilon)^\top \tag{335}$$

by Equation (333). Summing over neurons and using

$$\sum_{j \in \mathcal{J}_q^+} c_j = t_q, \qquad \sum_{j \in \mathcal{J}_q^-} c_j = 1 - t_q, \tag{336}$$

gives

$$\mathbf{H}^\top \mathbf{H} = \sum_{q \in \mathcal{E}} \|\boldsymbol{\beta}_q\| \big(t_q \mathbf{f}_q^+ (\mathbf{f}_q^+)^\top + (1-t_q) \mathbf{f}_q^- (\mathbf{f}_q^-)^\top \big). \tag{337}$$

This proves the first claim that every parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ attaining the lower bound in Equation (314) induces a split $\mathbf{t} \in \mathcal{T}_{N_h}$ for which its RSM is given by Equation (23).

For the converse, fix any $\mathbf{t} \in \mathcal{T}_{N_h}$. For each $q \in \mathcal{E}$, introduce neurons with parameters

$$\begin{aligned} \overline{\mathbf{w}}_{q,+} &= \sqrt{t_q \|\boldsymbol{\beta}_q\|} \, \overline{\mathbf{w}}_q, & \mathbf{a}_{q,+} &= \sqrt{\frac{t_q}{\|\boldsymbol{\beta}_q\|}} \, \boldsymbol{\beta}_q, \\ \overline{\mathbf{w}}_{q,-} &= -\sqrt{(1-t_q)\|\boldsymbol{\beta}_q\|} \, \overline{\mathbf{w}}_q, & \mathbf{a}_{q,-} &= \sqrt{\frac{1-t_q}{\|\boldsymbol{\beta}_q\|}} \, \boldsymbol{\beta}_q, \end{aligned} \tag{338}$$

omitting the positively oriented neuron when $t_q = 0$ and the negatively oriented neuron when $t_q = 1$. This construction uses exactly $\kappa(\mathbf{t})$ neurons. Since $\mathbf{t} \in \mathcal{T}_{N_h}$, and hence $\kappa(\mathbf{t}) \leq N_h$, adding $N_h - \kappa(\mathbf{t})$ neurons with all parameters set to zero yields a parameterization of width N_h. By construction, the positively and negatively oriented neurons contribute

$$t_q \boldsymbol{\beta}_q \qquad \text{and} \qquad (1-t_q)\boldsymbol{\beta}_q, \tag{339}$$

respectively, to the class aggregate $\boldsymbol{\beta}_q$. Moreover, $\mathbf{t} \in \mathcal{T}$ ensures that the resulting global residual equals $\mathbf{r}$. Since no nonessential parameter class is introduced, Proposition 3.3 therefore implies that the constructed parameterization belongs to $\mathcal{O}_{N_h}$.

It remains to show that the constructed parameterization has weight norm $2\sum_{q \in \mathcal{E}} \|\boldsymbol{\beta}_q\|$ and that its RSM is of the form in Equation (23). The former follows directly from the observation that the constructed parameterization satisfies all three conditions of Lemma F.1: neurons outside essential

parameter classes have parameters $\overline{\mathbf{w}}_j = \mathbf{0}$ and $\mathbf{a}_j = \mathbf{0}$, those in essential classes have equal incoming-parameter and readout norms,

$$\begin{aligned}\|\overline{\mathbf{w}}_{q,+}\| &= \sqrt{t_q\|\boldsymbol{\beta}_q\|} = \sqrt{\frac{t_q}{\|\boldsymbol{\beta}_q\|}}\,\|\boldsymbol{\beta}_q\| = \|\mathbf{a}_{q,+}\|,\\ \|\overline{\mathbf{w}}_{q,-}\| &= \sqrt{(1-t_q)\|\boldsymbol{\beta}_q\|} = \sqrt{\frac{1-t_q}{\|\boldsymbol{\beta}_q\|}}\,\|\boldsymbol{\beta}_q\| = \|\mathbf{a}_{q,-}\|,\end{aligned} \tag{340}$$

and, for every $q \in \mathcal{E}$,

$$\|\overline{\mathbf{w}}_{q,+}\|\mathbf{a}_{q,+} = t_q\boldsymbol{\beta}_q, \qquad \|\overline{\mathbf{w}}_{q,-}\|\mathbf{a}_{q,-} = (1-t_q)\boldsymbol{\beta}_q, \tag{341}$$

where $t_q \geq 0$, $1 - t_q \geq 0$, and $t_q + (1 - t_q) = 1$, with the corresponding term omitted when an orientation is absent. For the latter, positive 1-homogeneity gives feature rows

$$\sqrt{t_q\|\boldsymbol{\beta}_q\|}\,(\mathbf{f}_q^+)^\top \qquad \text{and} \qquad \sqrt{(1-t_q)\|\boldsymbol{\beta}_q\|}\,(\mathbf{f}_q^-)^\top, \tag{342}$$

so the constructed parameterization realizes the RSM in Equation (23) for the chosen $\mathbf{t}$. □

The preceding proof demonstrates that $\mathcal{T}_{N_h} \neq \emptyset$ is necessary and sufficient for the lower bound in Equation (314) to be attained. When this condition is satisfied, Proposition 6.1 characterizes the RSMs of norm-minimizing parameterizations, rather than the parameterizations themselves. For a fixed split $\mathbf{t}$, the coefficients $c_j \geq 0$ can be redistributed among neurons of either orientation while preserving their sums t_q and $1 - t_q$. Such redistributions leave the RSM unchanged, so distinct MWNPs can realize the same split $\mathbf{t}$ and thus the same RSM.

A separate degree of freedom lies in the split $\mathbf{t}$ itself, since different elements of $\mathcal{T}_{N_h}$ can yield different RSMs through Equation (23). Under the linear-independence condition of Corollary 6.2, the residual constraint in Equation (22) admits at most one split. Consequently, all MWNPs induce the same RSM.

Proof of Corollary 6.2. Let $\mathbf{t}, \mathbf{s} \in \mathcal{T}$ be two residual-compatible splits. Since $\mathbf{t}, \mathbf{s} \in \mathcal{T}$, both satisfy Equation (22). Hence, for every $\mathbf{x} \in \mathbb{R}^{N_i}$, the expressions corresponding to $\mathbf{t}$ and $\mathbf{s}$ are both equal to $\mathbf{r}(\mathbf{x})$. Taking their difference and using $m \neq 0$ gives

$$\left[\sum_{q\in\mathcal{E}}(t_q - s_q)\boldsymbol{\beta}_q\overline{\mathbf{w}}_q^\top\right]\overline{\mathbf{x}} = \mathbf{0}, \qquad \mathbf{x} \in \mathbb{R}^{N_i}. \tag{343}$$

Since this expression vanishes for every input $\mathbf{x} \in \mathbb{R}^{N_i}$, we have

$$\sum_{q\in\mathcal{E}}(t_q - s_q)\boldsymbol{\beta}_q\overline{\mathbf{w}}_q^\top = \mathbf{0}. \tag{344}$$

Linear independence of the matrices $\boldsymbol{\beta}_q\overline{\mathbf{w}}_q^\top$, $q \in \mathcal{E}$, therefore implies $\mathbf{t} = \mathbf{s}$. Hence, $\mathcal{T}$ contains at most one split. Since $\mathcal{T}_{N_h} \subseteq \mathcal{T}$ and $\mathcal{T}_{N_h} \neq \emptyset$, $\mathcal{T}$ contains a unique split $\mathbf{t}$, which belongs to $\mathcal{T}_{N_h}$. Proposition 6.1 therefore gives a unique minimum-weight-norm RSM.

For every $L \geq N_h$, this split remains feasible because $\kappa(\mathbf{t}) \leq N_h \leq L$ and hence belongs to $\mathcal{T}_L$. Since $\mathbf{t}$ is the unique element of $\mathcal{T}$, it is also the unique element of $\mathcal{T}_L$. Applying Proposition 6.1 again yields the same RSM. □

The preceding result guarantees identifiability by showing that the residual constraint admits at most one split. For even positively 1-homogeneous activations, neurons of either orientation generate identical features $\mathbf{f}_q^+ = \mathbf{f}_q^-$. In this case, every residual-compatible split $\mathbf{t} \in \mathcal{T}$ yields the same RSM, so representational identifiability holds irrespective of whether the residual constraint admits multiple splits.

Corollary F.3 (Identifiability for even positively 1-homogeneous activations)**.** *Suppose $\sigma(z) = \delta|z|$ with $\delta \neq 0$, and let $\mathcal{O}$ be a symmetry orbit with residual $\mathbf{r} = \mathbf{0}$. At every width $N_h \geq |\mathcal{E}|$, the minimum of Ω_W is attained, and all MWNPs have the same RSM, namely*

$$\sum_{q\in\mathcal{E}}\|\boldsymbol{\beta}_q\|\,\mathbf{f}_q^+(\mathbf{f}_q^+)^\top. \tag{345}$$

Proof. Since $m = 0$ and $\mathbf{r} = \mathbf{0}$, every $\mathbf{t} \in [0,1]^{|\mathcal{E}|}$ satisfies the residual constraint in Equation (22). Choosing $\mathbf{t} \in \{0,1\}^{|\mathcal{E}|}$ gives $\kappa(\mathbf{t}) = |\mathcal{E}|$, so $\mathcal{T}_{N_h} \neq \emptyset$ for every $N_h \geq |\mathcal{E}|$. Moreover, since σ is even, $\mathbf{f}_q^+ = \mathbf{f}_q^-$ for every $q \in \mathcal{E}$. Hence, Equation (23) reduces to

$$\sum_{q \in \mathcal{E}} \|\boldsymbol{\beta}_q\| \, \mathbf{f}_q^+ (\mathbf{f}_q^+)^\top, \tag{346}$$

independently of $\mathbf{t}$. The result therefore follows from Proposition 6.1. □

For positively 1-homogeneous activations with $m \neq 0$, Proposition 6.1 shows that minimum-weight-norm parameterizations need not have a unique RSM: when $\mathcal{T}_{N_h}$ contains multiple residual-compatible splits, Equation (23) parameterizes the corresponding RSMs, which can be distinct. Corollary 6.2 shows that such a family cannot arise when the matrices $\boldsymbol{\beta}_q \overline{\mathbf{w}}_q^\top$, $q \in \mathcal{E}$, are linearly independent, since the residual constraint then admits at most one split. Thus, whenever $\mathcal{T}_{N_h} \neq \emptyset$, any orbit admitting multiple distinct minimum-weight-norm RSMs must violate this linear-independence condition. The following ReLU example illustrates precisely this case.

Example F.4 (Nonunique minimum-weight-norm geometry)**.** Consider the ReLU tent function $f_{\boldsymbol{\theta}^+}(x) = \psi(1 - |x|)$ from Example B.13. Recall that the parameterization $\boldsymbol{\theta}^+$ has incoming parameters and readouts

$$(\overline{\mathbf{w}}_1, a_1) = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 1 \right), \qquad (\overline{\mathbf{w}}_2, a_2) = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, -2 \right), \qquad (\overline{\mathbf{w}}_3, a_3) = \left(\begin{bmatrix} 1 \\ -1 \end{bmatrix}, 1 \right). \tag{347}$$

The three essential parameter classes are

$$q_j = [\overline{\mathbf{w}}_j], \qquad j \in \{1,2,3\}. \tag{348}$$

Choose the unit-norm representatives

$$\overline{\mathbf{w}}_{q_1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \overline{\mathbf{w}}_{q_2} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \overline{\mathbf{w}}_{q_3} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{349}$$

Since each essential parameter class contains a single neuron of $\boldsymbol{\theta}^+$, the definition of the orbit invariants β_q in Equation (65) gives

$$(\beta_{q_1}, \beta_{q_2}, \beta_{q_3}) = \left(\sqrt{2}, -2, \sqrt{2}\right). \tag{350}$$

Hence,

$$\beta_{q_1} \overline{\mathbf{w}}_{q_1}^\top + \beta_{q_2} \overline{\mathbf{w}}_{q_2}^\top + \beta_{q_3} \overline{\mathbf{w}}_{q_3}^\top = (1,1) - 2(1,0) + (1,-1) = (0,0), \tag{351}$$

so the matrices $\beta_q \overline{\mathbf{w}}_q^\top$ are linearly dependent.

Since the orbit residual vanishes, the residual constraint in Equation (22) becomes

$$(2t_{q_1} - 1)(1,1) - 2(2t_{q_2} - 1)(1,0) + (2t_{q_3} - 1)(1,-1) = (0,0). \tag{352}$$

Equivalently,

$$t_{q_1} - 2t_{q_2} + t_{q_3} = 0 \tag{353}$$

and

$$t_{q_1} - t_{q_3} = 0. \tag{354}$$

Equation (354) gives $t_{q_1} = t_{q_3}$, and substituting this into Equation (353) gives $t_{q_1} = t_{q_2}$. Therefore,

$$\mathcal{T} = \{(t,t,t) \mid t \in [0,1]\}. \tag{355}$$

Thus, the residual constraint leaves exactly one degree of freedom in the orientation split. For the endpoint splits $t \in \{0,1\}$, each essential class uses only one orientation, so

$$\kappa(t,t,t) = 3. \tag{356}$$

For every interior split $0 < t < 1$, both orientations are required in all three classes, giving

$$\kappa(t,t,t) = 6. \tag{357}$$

Hence,

$$\mathcal{T}_{N_h} = \{(0,0,0), (1,1,1)\} \qquad \text{for } 3 \leq N_h < 6, \tag{358}$$

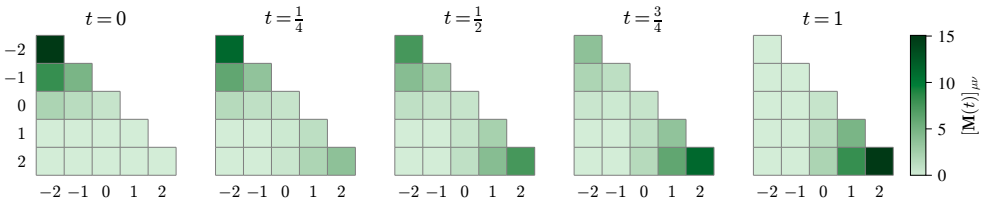

Figure F.1: **Nonunique minimum-weight-norm geometries of the ReLU tent function**. The one-parameter family of RSMs $\mathbf{M}(t) = t\mathbf{M}_+ + (1-t)\mathbf{M}_-$ realized by minimum-weight-norm parameterizations of the ReLU tent function $\psi(1-|x|)$ from Example F.4, evaluated on the inputs collected in $\mathbf{X} = (-2,-1,0,1,2)$. Each panel shows $\mathbf{M}(t)$ for one value of $t \in [0,1]$, corresponding to the residual-compatible split $\mathbf{t} = (t,t,t)$, with rows and columns indexed by the inputs and all panels sharing a common color scale. The endpoints $t=0$ and $t=1$ use a single orientation per essential class and correspond to the oppositely oriented parameterizations $\boldsymbol{\theta}^-$ and $\boldsymbol{\theta}^+$ of Figure B.1, rescaled to minimum-weight-norm parameterizations. Their RSMs are mirror images because reversing the inputs swaps the two orientations. Every interior value $0 < t < 1$ requires both orientations in all three essential classes. The diagonal entry corresponding to $x = 0$ is constant across the family, whereas every other nonzero entry varies with t, so distinct values of t yield distinct RSMs.

whereas $\mathcal{T}_{N_h} = \mathcal{T}$ for every $N_h \geq 6$.

By Proposition 6.1, every $\mathbf{t} \in \mathcal{T}_{N_h}$ is realized by an MWNP, and all such parameterizations attain the minimum weight norm

$$2\sum_{q\in\mathcal{E}}|\beta_q| = 4 + 4\sqrt{2}. \tag{359}$$

For any split $\mathbf{t} = (t,t,t) \in \mathcal{T}$, Equation (23) gives the RSM

$$\mathbf{M}(t) \coloneqq t\mathbf{M}_+ + (1-t)\mathbf{M}_-, \tag{360}$$

where

$$\mathbf{M}_+ \coloneqq \sum_{q\in\mathcal{E}}|\beta_q|\,\mathbf{f}_q^+(\mathbf{f}_q^+)^\top, \qquad \mathbf{M}_- \coloneqq \sum_{q\in\mathcal{E}}|\beta_q|\,\mathbf{f}_q^-(\mathbf{f}_q^-)^\top. \tag{361}$$

For the chosen unit-norm representatives, the two feature orientations $\mathbf{f}_q^+$ and $\mathbf{f}_q^-$ have entries

$$\begin{aligned}
(\mathbf{f}_{q_1}^+)_\mu &= \frac{1}{\sqrt{2}}\,\psi(x^\mu+1), & (\mathbf{f}_{q_1}^-)_\mu &= \frac{1}{\sqrt{2}}\,\psi(-x^\mu-1), \\
(\mathbf{f}_{q_2}^+)_\mu &= \psi(x^\mu), & (\mathbf{f}_{q_2}^-)_\mu &= \psi(-x^\mu), \qquad \mu = 1,\dots,P. \\
(\mathbf{f}_{q_3}^+)_\mu &= \frac{1}{\sqrt{2}}\,\psi(x^\mu-1), & (\mathbf{f}_{q_3}^-)_\mu &= \frac{1}{\sqrt{2}}\,\psi(1-x^\mu),
\end{aligned} \tag{362}$$

The geometries $\mathbf{M}_+$ and $\mathbf{M}_-$ in Equation (361), corresponding to the endpoints $t=1$ and $t=0$, respectively, are distinct on any input set containing an input $x^\mu > 1$. At such an input,

$$(\mathbf{f}_q^+)_\mu > 0, \qquad (\mathbf{f}_q^-)_\mu = 0 \qquad \text{for every } q \in \mathcal{E}, \tag{363}$$

so $\mathbf{M}_+ \neq \mathbf{M}_-$.

On such an input set, distinct values of t yield distinct RSMs, since

$$\mathbf{M}(t) - \mathbf{M}(s) = (t-s)(\mathbf{M}_+ - \mathbf{M}_-) \neq \mathbf{0} \tag{364}$$

whenever $t \neq s$. Thus, minimum-weight-norm selection yields two distinct RSMs for $3 \leq N_h < 6$ and a one-parameter family of distinct RSMs for every $N_h \geq 6$, visualized in Figure F.1.

The preceding example shows that minimum-weight-norm selection need not identify a unique representational geometry when the linear-independence condition of Corollary 6.2 fails. We close this analysis by returning to the six analytical ReLU solutions underlying Figure 1, for which the main text asserts representational identifiability under minimum-weight-norm selection. For each solution, we verify that the conditions of Corollary 6.2 hold, so the selected RSM is unique and remains unchanged under further increases in width.

Example F.5 (Identifiability of the analytical ReLU solutions)**.** Consider the six analytical ReLU solutions illustrated in Figure 1 and detailed in Table G.1. Following the convention described in Appendix G.1, we hold the output bias b fixed and apply Corollary 6.2 to $f_{\boldsymbol{\theta}} - b$.

For solutions 1, 3, 4, and 6, the two incoming-parameter vectors have the form

$$\overline{\mathbf{w}}_1 = \begin{bmatrix} \mathbf{v} \\ 0 \end{bmatrix} \qquad \text{and} \qquad \overline{\mathbf{w}}_2 = \begin{bmatrix} \mathbf{v} \\ \sqrt{2} \end{bmatrix}, \qquad \|\mathbf{v}\| = 1, \tag{365}$$

up to ordering. These vectors are linearly independent and therefore belong to distinct parameter classes

$$q_1 = [\overline{\mathbf{w}}_1], \qquad q_2 = [\overline{\mathbf{w}}_2], \tag{366}$$

both of which are essential because their corresponding readouts a_1 and a_2 are nonzero. Choose the unit-norm representatives

$$\overline{\mathbf{w}}_{q_j} = \frac{\overline{\mathbf{w}}_j}{\|\overline{\mathbf{w}}_j\|}, \qquad j \in \{1, 2\}. \tag{367}$$

Since each class contains a single neuron of the given solution, Equation (65) gives

$$\beta_{q_j} = \|\overline{\mathbf{w}}_j\| a_j, \qquad \beta_{q_j} \overline{\mathbf{w}}_{q_j}^\top = a_j \overline{\mathbf{w}}_j^\top, \qquad j \in \{1, 2\}. \tag{368}$$

Because $a_1, a_2 \neq 0$ and $\overline{\mathbf{w}}_1, \overline{\mathbf{w}}_2$ are linearly independent, the two matrices $\beta_{q_1} \overline{\mathbf{w}}_{q_1}^\top$ and $\beta_{q_2} \overline{\mathbf{w}}_{q_2}^\top$ are linearly independent. Moreover, the two neurons realize the split $\mathbf{t} = (1, 1)$, so $\mathbf{t} \in \mathcal{T}_2$ and hence $\mathcal{T}_2 \neq \emptyset$.

For solutions 2 and 5, the two incoming-parameter vectors are opposite unit vectors with zero hidden bias and therefore belong to a single essential parameter class. Their readout weights are $(1, 1)$ and $(-1, -1)$, respectively, giving the scalar class aggregate $\beta_q = 2$ for solution 2 and $\beta_q = -2$ for solution 5. In both cases, the residual vanishes, so Equation (22) uniquely fixes $t_q = 1/2$. Since $\kappa(t_q) = 2$, this split belongs to $\mathcal{T}_2$. Moreover, $\beta_q \overline{\mathbf{w}}_q^\top \neq \mathbf{0}$, so the singleton family $\{\beta_q \overline{\mathbf{w}}_q^\top\}$ is linearly independent.

Since ReLU has $m = 1/2 \neq 0$, all six solutions thus satisfy the assumptions of Corollary 6.2 at width $N_h = 2$. Each corresponding symmetry orbit therefore has a unique minimum-weight-norm RSM, which remains unchanged at every width $N_h \geq 2$.

F.2 Minimum representation norm for positively 1-homogeneous activations

We now turn to minimum-representation-norm parameterizations, whose analysis closely parallels that of the minimum-weight-norm case presented in the preceding subsection. We first derive an orbit-wide lower bound on the representation-norm objective Ω_H and characterize exactly when this bound is attained. The resulting equality conditions determine the admissible orientation splits and yield the corresponding family of minimum-representation-norm RSMs, analogous to Proposition 6.1. We then identify conditions under which this family reduces to a unique geometry, similar to Corollary 6.2. Finally, in contrast to the MWNP case, we show that a fixed symmetry orbit need not contain an MRNP at all.

Throughout this subsection, we retain the setting and notation of Appendix F.1, including the symmetry orbit $\mathcal{O}$, essential parameter classes $\mathcal{E}$, orbit invariants $\boldsymbol{\beta}_q$, residual $\mathbf{r}$, unit-norm representatives $\overline{\mathbf{w}}_q$, feature vectors $\mathbf{f}_q^+$ and $\mathbf{f}_q^-$, and residual-compatible split set $\mathcal{T}$.

Unlike the weight-norm objective Ω_W, the representation-norm objective Ω_H depends on the fixed input matrix $\mathbf{X}$ through the rows $\mathbf{v}_j^\top = \sigma(\overline{\mathbf{w}}_j^\top \overline{\mathbf{X}})$ of the hidden-activation matrix $\mathbf{H}$, since

$$\|\mathbf{H}\|_F^2 = \sum_{j=1}^{N_h} \|\mathbf{v}_j\|^2. \tag{369}$$

For each essential parameter class $q \in \mathcal{E}$, define

$$g_q^+ := \|\mathbf{f}_q^+\|, \qquad g_q^- := \|\mathbf{f}_q^-\|, \qquad g_q := \min\{g_q^+, g_q^-\}. \tag{370}$$

The next lemma is the minimum-representation-norm analog of Lemma F.1. It again gives an orbit-wide lower bound together with necessary and sufficient equality conditions. Here, achieving the

lower bound requires matching the magnitude $\|\mathbf{v}_j\|$ of the feature computed by the jth neuron to the magnitude $\|\mathbf{a}_j\|$ of its readout. For any essential class $q \in \mathcal{E}$ with $g_q^+ \neq g_q^-$, achieving the lower bound further requires every neuron in that class to have the orientation attaining $g_q = \min\{g_q^+, g_q^-\}$.

Lemma F.6 (Representation-norm lower bound and equality conditions)**.** *Suppose $g_q > 0$ for every essential parameter class $q \in \mathcal{E}$. Then, every parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ satisfies*

$$\Omega_H(\boldsymbol{\theta}) \geq 2 \sum_{q \in \mathcal{E}} g_q \|\boldsymbol{\beta}_q\|. \tag{371}$$

Equality holds if and only if the following three conditions are satisfied:

(i) Every neuron outside the essential parameter classes has

$$\mathbf{v}_j = \mathbf{0}, \qquad \mathbf{a}_j = \mathbf{0}. \tag{372}$$

(ii) Every neuron in an essential parameter class has equal feature and readout norms:

$$\|\mathbf{v}_j\| = \|\mathbf{a}_j\|, \qquad j \in \mathcal{J}_q, \quad q \in \mathcal{E}. \tag{373}$$

Moreover, whenever the two orientations of a class $q \in \mathcal{E}$ have different feature norms $g_q^+ \neq g_q^-$ on $\mathbf{X}$, all neurons in q have the orientation with the smaller feature norm. Equivalently,

$$g_q^+ < g_q^- \implies \mathcal{J}_q^- = \emptyset, \qquad g_q^- < g_q^+ \implies \mathcal{J}_q^+ = \emptyset. \tag{374}$$

(iii) For every $q \in \mathcal{E}$, there exist coefficients $c_j \geq 0$, $j \in \mathcal{J}_q$, such that

$$\|\overline{\mathbf{w}}_j\|\mathbf{a}_j = c_j \boldsymbol{\beta}_q, \qquad \sum_{j \in \mathcal{J}_q} c_j = 1. \tag{375}$$

Proof. By definition of the representation-norm objective $\Omega_H(\boldsymbol{\theta})$ in Equation (18), we have

$$\Omega_H(\boldsymbol{\theta}) = \sum_{j=1}^{N_h} \big(\|\mathbf{v}_j\|^2 + \|\mathbf{a}_j\|^2\big) \geq \sum_{q \in \mathcal{E}} \sum_{j \in \mathcal{J}_q} \big(\|\mathbf{v}_j\|^2 + \|\mathbf{a}_j\|^2\big). \tag{376}$$

The AM–GM inequality gives

$$\sum_{q \in \mathcal{E}} \sum_{j \in \mathcal{J}_q} \big(\|\mathbf{v}_j\|^2 + \|\mathbf{a}_j\|^2\big) \geq 2 \sum_{q \in \mathcal{E}} \sum_{j \in \mathcal{J}_q} \|\mathbf{v}_j\| \|\mathbf{a}_j\|. \tag{377}$$

For each essential parameter class $q \in \mathcal{E}$, positive 1-homogeneity and the unit norm of $\overline{\mathbf{w}}_q$ give

$$\|\mathbf{v}_j\| = \begin{cases} \|\overline{\mathbf{w}}_j\| g_q^+, & j \in \mathcal{J}_q^+, \\ \|\overline{\mathbf{w}}_j\| g_q^-, & j \in \mathcal{J}_q^-. \end{cases} \tag{378}$$

Since $g_q = \min\{g_q^+, g_q^-\}$, it follows that

$$\begin{aligned} \sum_{j \in \mathcal{J}_q} \|\mathbf{v}_j\| \|\mathbf{a}_j\| &= g_q^+ \sum_{j \in \mathcal{J}_q^+} \|\overline{\mathbf{w}}_j\| \|\mathbf{a}_j\| + g_q^- \sum_{j \in \mathcal{J}_q^-} \|\overline{\mathbf{w}}_j\| \|\mathbf{a}_j\| \\ &\geq g_q \sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\| \|\mathbf{a}_j\|. \end{aligned} \tag{379}$$

Finally, absolute homogeneity of the Euclidean norm and the triangle inequality yield

$$\sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\| \|\mathbf{a}_j\| = \sum_{j \in \mathcal{J}_q} \big\| \|\overline{\mathbf{w}}_j\| \mathbf{a}_j \big\| \geq \Big\| \sum_{j \in \mathcal{J}_q} \|\overline{\mathbf{w}}_j\| \mathbf{a}_j \Big\| = \|\boldsymbol{\beta}_q\|, \tag{380}$$

where the final equality follows from the definition of the class aggregate $\boldsymbol{\beta}_q$ in Equation (65). Together, Equations (376), (377), (379), and (380) establish Equation (371).

Equality in Equation (371) holds if and only if equality holds in Equations (376) and (377) and, for every $q \in \mathcal{E}$, in Equations (379) and (380). It therefore remains to determine when each inequality is tight.

Equality in Equation (376) holds if and only if every neuron outside the essential parameter classes contributes nothing to Ω_H. Since each such contribution is

$$\|\mathbf{v}_j\|^2 + \|\mathbf{a}_j\|^2 \geq 0, \tag{381}$$

this is equivalent to

$$\mathbf{v}_j = \mathbf{0}, \qquad \mathbf{a}_j = \mathbf{0}, \tag{382}$$

for every neuron outside the essential parameter classes, giving the first condition.

Equality in Equation (377) holds if and only if the AM–GM inequality is tight for every neuron in every essential parameter class. For each $j \in \mathcal{J}_q$,

$$\left(\|\mathbf{v}_j\|^2 + \|\mathbf{a}_j\|^2\right) - 2\|\mathbf{v}_j\|\|\mathbf{a}_j\| = \left(\|\mathbf{v}_j\| - \|\mathbf{a}_j\|\right)^2, \tag{383}$$

so equality holds if and only if

$$\|\mathbf{v}_j\| = \|\mathbf{a}_j\|, \qquad j \in \mathcal{J}_q, \quad q \in \mathcal{E}. \tag{384}$$

Let $\varepsilon \in \{+, -\}$. Because $g_q^\varepsilon \geq g_q > 0$ and every neuron in $\mathcal{J}_q^\varepsilon$ has $\|\overline{\mathbf{w}}_j\| > 0$, this equality also implies $\|\mathbf{a}_j\| > 0$ for every neuron in an essential parameter class. Consequently, equality in Equation (379) holds if and only if every represented orientation of class q has feature norm g_q. Equivalently,

$$g_q^+ < g_q^- \implies \mathcal{J}_q^- = \emptyset, \qquad g_q^- < g_q^+ \implies \mathcal{J}_q^+ = \emptyset, \tag{385}$$

completing the second condition.

Finally, equality in Equation (380) is determined by equality in the triangle inequality. Since $\boldsymbol{\beta}_q \neq \mathbf{0}$ for every essential parameter class $q \in \mathcal{E}$, equality holds if and only if the vectors

$$\|\overline{\mathbf{w}}_j\|\mathbf{a}_j, \qquad j \in \mathcal{J}_q, \tag{386}$$

are nonnegative multiples of $\boldsymbol{\beta}_q$. Equivalently, there exist coefficients $c_j \geq 0$ such that

$$\|\overline{\mathbf{w}}_j\|\mathbf{a}_j = c_j \boldsymbol{\beta}_q, \qquad j \in \mathcal{J}_q. \tag{387}$$

Summing over $j \in \mathcal{J}_q$ and using the definition of the class aggregate gives

$$\boldsymbol{\beta}_q = \boldsymbol{\beta}_q \sum_{j \in \mathcal{J}_q} c_j. \tag{388}$$

Since $\boldsymbol{\beta}_q \neq \mathbf{0}$, we have $\sum_{j \in \mathcal{J}_q} c_j = 1$, giving the third condition.

Conversely, if the three stated conditions hold, then equality holds in Equations (376), (377), (379), and (380), and hence in Equation (371). □

As in the minimum-weight-norm case, the equality conditions in Lemma F.6 again determine how the class aggregate $\boldsymbol{\beta}_q$ can be distributed between the two orientations of each essential parameter class at the representation-norm lower bound.

Corollary F.7 (Readout allocation at the representation-norm lower bound)**.** *Suppose that $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ attains the lower bound in Equation (371). Then, for every essential parameter class $q \in \mathcal{E}$, there exists $t_q \in [0, 1]$ such that*

$$\sum_{j \in \mathcal{J}_q^+} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = t_q \boldsymbol{\beta}_q, \qquad \sum_{j \in \mathcal{J}_q^-} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = (1 - t_q)\boldsymbol{\beta}_q. \tag{389}$$

Moreover,

$$g_q^+ < g_q^- \implies t_q = 1, \qquad g_q^- < g_q^+ \implies t_q = 0. \tag{390}$$

Proof. By Lemma F.6, there exist coefficients $c_j \geq 0$, $j \in \mathcal{J}_q$, such that

$$\|\overline{\mathbf{w}}_j\|\mathbf{a}_j = c_j \boldsymbol{\beta}_q, \qquad \sum_{j \in \mathcal{J}_q} c_j = 1. \tag{391}$$

Setting

$$t_q \coloneqq \sum_{j \in \mathcal{J}_q^+} c_j \tag{392}$$

gives $t_q \in [0, 1]$ and $\sum_{j \in \mathcal{J}_q^-} c_j = 1 - t_q$, yielding the two allocation identities. If $g_q^+ < g_q^-$, Lemma F.6 gives $\mathcal{J}_q^- = \emptyset$, so $t_q = 1$. Similarly, if $g_q^- < g_q^+$, then $\mathcal{J}_q^+ = \emptyset$ and hence $t_q = 0$. □

The preceding corollary shows that every parameterization attaining the representation-norm lower bound $2\sum_{q\in\mathcal{E}} g_q\|\boldsymbol{\beta}_q\|$ induces a residual-compatible split $\mathbf{t}$ whose coordinates t_q are additionally constrained whenever the two features $\mathbf{f}_q^+$ and $\mathbf{f}_q^-$ of an essential class q have different magnitudes. Accordingly, let $\mathcal{T}^H \subseteq \mathcal{T}$ denote the residual-compatible splits $\mathbf{t} \in \mathcal{T}$ satisfying, for every $q \in \mathcal{E}$,

$$t_q = \begin{cases} 1, & \text{if } g_q^+ < g_q^-, \\ 0, & \text{if } g_q^- < g_q^+, \end{cases} \tag{393}$$

with no additional restriction imposed on t_q when $g_q^+ = g_q^-$. As in the minimum-weight-norm case, we further restrict to splits realizable at width N_h by defining

$$\mathcal{T}_{N_h}^H := \{\mathbf{t} \in \mathcal{T}^H \mid \kappa(\mathbf{t}) \le N_h\} = \mathcal{T}^H \cap \mathcal{T}_{N_h}. \tag{394}$$

Thus, $\mathcal{T}_{N_h}^H$ contains precisely the residual-compatible splits $\mathbf{t}$ that can be realized at width N_h using only orientations permitted by the equality conditions of Lemma F.6.

We can now characterize the RSMs of all parameterizations attaining the representation-norm lower bound, in direct analogy with Proposition 6.1.

Proposition F.8 (Minimum-representation-norm representational geometry)**.** *Fix a symmetry orbit $\mathcal{O}$ with positively 1-homogeneous activation and essential parameter classes $\mathcal{E}$, and suppose $g_q > 0$ for every $q \in \mathcal{E}$. If $\mathcal{T}_{N_h}^H \neq \emptyset$, the minimum of Ω_H over $\mathcal{O}_{N_h}$ is attained and equals*

$$2\sum_{q\in\mathcal{E}} g_q\|\boldsymbol{\beta}_q\|. \tag{395}$$

Moreover, the RSMs of all MRNPs in $\mathcal{O}_{N_h}$ are precisely

$$\sum_{q\in\mathcal{E}} \frac{\|\boldsymbol{\beta}_q\|}{g_q}\left(t_q\mathbf{f}_q^+(\mathbf{f}_q^+)^\top + (1-t_q)\mathbf{f}_q^-(\mathbf{f}_q^-)^\top\right), \tag{396}$$

as $\mathbf{t}$ ranges over $\mathcal{T}_{N_h}^H$.

Proof. We first show that every parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ attaining the lower bound in Equation (371) induces a split $\mathbf{t} \in \mathcal{T}_{N_h}^H$ for which its RSM is given by Equation (396).

Fix $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ attaining the lower bound in Equation (371). By Corollary F.7, for every $q \in \mathcal{E}$, there exists $t_q \in [0,1]$ such that

$$\sum_{j\in\mathcal{J}_q^+} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = t_q\boldsymbol{\beta}_q, \qquad \sum_{j\in\mathcal{J}_q^-} \|\overline{\mathbf{w}}_j\|\mathbf{a}_j = (1-t_q)\boldsymbol{\beta}_q. \tag{397}$$

By the first condition in Lemma F.6, every neuron outside the essential parameter classes has $\mathbf{a}_j = \mathbf{0}$ and therefore does not contribute to the global residual. Hence,

$$\mathbf{r}(\mathbf{x}) = m\sum_{q\in\mathcal{E}} (2t_q - 1)\boldsymbol{\beta}_q\overline{\mathbf{w}}_q^\top\overline{\mathbf{x}}. \tag{398}$$

Since $\boldsymbol{\theta} \in \mathcal{O}_{N_h}$ has residual $\mathbf{r}$, the definition of $\mathcal{T}$ in Equation (22) gives $\mathbf{t} \in \mathcal{T}$. Moreover, Corollary F.7 ensures that $\mathbf{t}$ satisfies the orientation restrictions in Equation (393), so $\mathbf{t} \in \mathcal{T}^H$. Since $\mathbf{t}$ is induced by a parameterization of width N_h, we have $\kappa(\mathbf{t}) \le N_h$ and therefore $\mathbf{t} \in \mathcal{T}_{N_h}^H$.

We next determine the RSM induced by $\boldsymbol{\theta}$. By the second condition in Lemma F.6, every neuron $j \in \mathcal{J}_q$ in every essential parameter class $q \in \mathcal{E}$ satisfies $\|\overline{\mathbf{w}}_j\|g_q = \|\mathbf{v}_j\| = \|\mathbf{a}_j\|$. By the third condition, there exist coefficients $c_j \ge 0$, $j \in \mathcal{J}_q$, such that $\|\overline{\mathbf{w}}_j\|\mathbf{a}_j = c_j\boldsymbol{\beta}_q$. Combining these gives

$$\|\overline{\mathbf{w}}_j\|^2 g_q = \|\overline{\mathbf{w}}_j\|\|\mathbf{a}_j\| = \big\|\|\overline{\mathbf{w}}_j\|\mathbf{a}_j\big\| = c_j\|\boldsymbol{\beta}_q\|. \tag{399}$$

Equivalently,

$$\|\overline{\mathbf{w}}_j\|^2 = \frac{c_j\|\boldsymbol{\beta}_q\|}{g_q}. \tag{400}$$

Let $\varepsilon \in \{+,-\}$. Positive 1-homogeneity implies that neuron $j \in \mathcal{J}_q^\varepsilon$ has feature row

$$\mathbf{v}_j^\top = \sigma(\overline{\mathbf{w}}_j^\top\overline{\mathbf{X}}) = \|\overline{\mathbf{w}}_j\|\sigma(\varepsilon\overline{\mathbf{w}}_q^\top\overline{\mathbf{X}}) = \|\overline{\mathbf{w}}_j\|(\mathbf{f}_q^\varepsilon)^\top. \tag{401}$$

Its contribution to $\mathbf{H}^\top\mathbf{H}$ is therefore

$$\mathbf{v}_j\mathbf{v}_j^\top = c_j\frac{\|\boldsymbol{\beta}_q\|}{g_q}\mathbf{f}_q^\varepsilon(\mathbf{f}_q^\varepsilon)^\top \tag{402}$$

by Equation (400). Summing over neurons and using

$$\sum_{j\in\mathcal{J}_q^+} c_j = t_q, \qquad \sum_{j\in\mathcal{J}_q^-} c_j = 1-t_q, \tag{403}$$

gives

$$\mathbf{H}^\top\mathbf{H} = \sum_{q\in\mathcal{E}}\frac{\|\boldsymbol{\beta}_q\|}{g_q}\big(t_q\mathbf{f}_q^+(\mathbf{f}_q^+)^\top + (1-t_q)\mathbf{f}_q^-(\mathbf{f}_q^-)^\top\big). \tag{404}$$

This proves the first claim that every parameterization $\boldsymbol{\theta}\in\mathcal{O}_{N_h}$ attaining the lower bound in Equation (371) induces a split $\mathbf{t}\in\mathcal{T}_{N_h}^H$ for which its RSM is given by Equation (396).

For the converse, fix any $\mathbf{t}\in\mathcal{T}_{N_h}^H$. For each $q\in\mathcal{E}$, introduce neurons with parameters

$$\begin{aligned}
\overline{\mathbf{w}}_{q,+} &= \sqrt{\frac{t_q\|\boldsymbol{\beta}_q\|}{g_q}}\,\overline{\mathbf{w}}_q, & \mathbf{a}_{q,+} &= \sqrt{\frac{t_q g_q}{\|\boldsymbol{\beta}_q\|}}\,\boldsymbol{\beta}_q,\\
\overline{\mathbf{w}}_{q,-} &= -\sqrt{\frac{(1-t_q)\|\boldsymbol{\beta}_q\|}{g_q}}\,\overline{\mathbf{w}}_q, & \mathbf{a}_{q,-} &= \sqrt{\frac{(1-t_q)g_q}{\|\boldsymbol{\beta}_q\|}}\,\boldsymbol{\beta}_q,
\end{aligned} \tag{405}$$

omitting the positively oriented neuron when $t_q=0$ and the negatively oriented neuron when $t_q=1$. This construction uses exactly $\kappa(\mathbf{t})$ neurons. Since $\mathbf{t}\in\mathcal{T}_{N_h}^H$, and hence $\kappa(\mathbf{t})\le N_h$, adding $N_h-\kappa(\mathbf{t})$ neurons with all parameters set to zero yields a parameterization of width N_h.

By construction,

$$\|\overline{\mathbf{w}}_{q,+}\|\mathbf{a}_{q,+} = t_q\boldsymbol{\beta}_q, \qquad \|\overline{\mathbf{w}}_{q,-}\|\mathbf{a}_{q,-} = (1-t_q)\boldsymbol{\beta}_q. \tag{406}$$

Thus, the two orientations together reproduce the class aggregate $\boldsymbol{\beta}_q$. Moreover, $\mathbf{t}\in\mathcal{T}$ ensures that the resulting global residual equals $\mathbf{r}$. Since no nonessential parameter class is introduced, Proposition 3.3 therefore implies that the constructed parameterization belongs to $\mathcal{O}_{N_h}$.

It remains to show that the constructed parameterization has representation norm $2\sum_{q\in\mathcal{E}} g_q\|\boldsymbol{\beta}_q\|$ and that its RSM is of the form in Equation (396). The former follows by verifying the three conditions of Lemma F.6. The padding neurons have $\mathbf{v}_j=\mathbf{0}$ and $\mathbf{a}_j=\mathbf{0}$. Since $\mathbf{t}\in\mathcal{T}^H$, every represented orientation of class q has feature norm g_q on $\mathbf{X}$. Hence, the feature and readout norms of the constructed essential neurons satisfy

$$\begin{aligned}
\|\mathbf{v}_{q,+}\| &= \sqrt{\frac{t_q\|\boldsymbol{\beta}_q\|}{g_q}}\,g_q = \sqrt{t_q g_q\|\boldsymbol{\beta}_q\|} = \|\mathbf{a}_{q,+}\|,\\
\|\mathbf{v}_{q,-}\| &= \sqrt{\frac{(1-t_q)\|\boldsymbol{\beta}_q\|}{g_q}}\,g_q = \sqrt{(1-t_q)g_q\|\boldsymbol{\beta}_q\|} = \|\mathbf{a}_{q,-}\|.
\end{aligned} \tag{407}$$

Together with

$$\|\overline{\mathbf{w}}_{q,+}\|\mathbf{a}_{q,+} = t_q\boldsymbol{\beta}_q, \qquad \|\overline{\mathbf{w}}_{q,-}\|\mathbf{a}_{q,-} = (1-t_q)\boldsymbol{\beta}_q, \tag{408}$$

where $t_q\ge 0$, $1-t_q\ge 0$, and $t_q+(1-t_q)=1$, with the corresponding term omitted when an orientation is absent, all three equality conditions are satisfied. Thus, the constructed parameterization does indeed attain the lower bound in Equation (371).

For the RSM characterization, positive 1-homogeneity gives feature rows

$$\sqrt{\frac{t_q\|\boldsymbol{\beta}_q\|}{g_q}}\,(\mathbf{f}_q^+)^\top \qquad \text{and} \qquad \sqrt{\frac{(1-t_q)\|\boldsymbol{\beta}_q\|}{g_q}}\,(\mathbf{f}_q^-)^\top, \tag{409}$$

so the constructed parameterization realizes the RSM in Equation (396) for the chosen $\mathbf{t}$. □

Under the positivity assumption $g_q > 0$, the preceding proof shows that $\mathcal{T}^H_{N_h} \neq \emptyset$ is necessary and sufficient for the lower bound in Equation (371) to be attained. As in the minimum-weight-norm case, the resulting RSMs contain no auxiliary contributions. Unlike in that case, however, neurons from nonessential classes need not have zero incoming parameters. Any such neuron with $\mathbf{v}_j = \mathbf{0}$ and $\mathbf{a}_j = \mathbf{0}$ has zero representation-norm cost regardless of its incoming parameters.

Proposition F.8 further shows that minimum-representation-norm selection fixes the weights of essential feature contributions jointly through the orbit invariants $\boldsymbol{\beta}_q$ and the feature norms g_q on $\mathbf{X}$. The only remaining freedom is the residual-compatible allocation between orientations of equal feature norm $g_q^+ = g_q^-$. This freedom disappears either when one orientation has strictly smaller feature norm for every essential class or when the residual constraint uniquely determines the split.

Corollary F.9 (Minimum-representation-norm identifiability)**.** *In the setting of Proposition F.8, suppose additionally that either*

(i) $g_q^+ \neq g_q^-$ *for every* $q \in \mathcal{E}$*, or*

(ii) $m \neq 0$ *and the matrices* $\boldsymbol{\beta}_q \overline{\mathbf{w}}_q^\top$, $q \in \mathcal{E}$*, are linearly independent.*

Then, all MRNPs in $\mathcal{O}$ *of width at least* N_h *induce the same RSM.*

Proof. Under the first condition, Equation (393) uniquely fixes every coordinate t_q, so $\mathcal{T}^H$ contains at most one split. Under the second condition, the argument in the proof of Corollary 6.2 shows that $\mathcal{T}$, and hence $\mathcal{T}^H \subseteq \mathcal{T}$, contains at most one split. In either case, $\mathcal{T}^H_{N_h} \neq \emptyset$ guarantees that this unique split exists and is feasible at width N_h. Proposition F.8 therefore gives a unique minimum-representation-norm RSM. The argument in the proof of Corollary 6.2 then applies identically to show that this split remains feasible and unique at every larger width, yielding the same RSM. □

The preceding result guarantees identifiability by ensuring that $\mathcal{T}^H$ contains at most one split. For even positively 1-homogeneous activations, neurons of either orientation generate identical features $\mathbf{f}_q^+ = \mathbf{f}_q^-$ and hence satisfy $g_q^+ = g_q^- = g_q$. In this case, every split $\mathbf{t} \in \mathcal{T}^H_{N_h}$ yields the same RSM, so representational identifiability holds irrespective of whether $\mathcal{T}^H$ contains multiple splits.

Corollary F.10 (Identifiability for even positively 1-homogeneous activations)**.** *Suppose* $\sigma(z) = \delta|z|$ *with* $\delta \neq 0$*, and let* $\mathcal{O}$ *be a symmetry orbit with residual* $\mathbf{r} = \mathbf{0}$*. If* $g_q > 0$ *for every* $q \in \mathcal{E}$*, then at every width* $N_h \geq |\mathcal{E}|$*, the minimum of* Ω_H *is attained, and all MRNPs have the same RSM, namely*

$$\sum_{q \in \mathcal{E}} \frac{\|\boldsymbol{\beta}_q\|}{g_q} \mathbf{f}_q^+ (\mathbf{f}_q^+)^\top. \tag{410}$$

Proof. Since $m = 0$ and $\mathbf{r} = \mathbf{0}$, every $\mathbf{t} \in [0,1]^{|\mathcal{E}|}$ satisfies the residual constraint in Equation (22). Choosing $\mathbf{t} \in \{0,1\}^{|\mathcal{E}|}$ gives $\kappa(\mathbf{t}) = |\mathcal{E}|$, so $\mathcal{T}_{N_h} \neq \emptyset$ for every $N_h \geq |\mathcal{E}|$. Moreover, since σ is even, $\mathbf{f}_q^+ = \mathbf{f}_q^-$ and thus $g_q^+ = g_q^- = g_q$ for every $q \in \mathcal{E}$. Hence, $\mathcal{T}^H_{N_h} = \mathcal{T}_{N_h} \neq \emptyset$ for every $N_h \geq |\mathcal{E}|$, while $\mathbf{f}_q^+ = \mathbf{f}_q^-$ reduces Equation (396) to

$$\sum_{q \in \mathcal{E}} \frac{\|\boldsymbol{\beta}_q\|}{g_q} \mathbf{f}_q^+ (\mathbf{f}_q^+)^\top, \tag{411}$$

independently of $\mathbf{t}$. The result therefore follows from Proposition F.8. □

The positivity assumption $g_q > 0$ in Proposition F.8 excludes essential parameter classes for which one of the features $\mathbf{f}_q^+$ or $\mathbf{f}_q^-$ vanishes. The following example shows why this assumption matters: when an orbit $\mathcal{O}$ contains such an essential parameter class, the minimum of the representation-norm objective Ω_H over $\mathcal{O}$ may fail to exist.

Example F.11 (Nonexistence of an MRNP)**.** Let $\psi(z) := \max(0, z)$ and consider the family of symmetry-equivalent width-1 parameterizations

$$\boldsymbol{\theta}_\alpha = (\psi; \alpha, -2\alpha, \alpha^{-1}), \qquad \alpha > 0, \tag{412}$$

which realize $f_{\boldsymbol{\theta}_\alpha}(x) = \psi(x-2)$. On the inputs collected in $\mathbf{X} = (-1, 0, 1)$, the hidden-activation matrix of each parameterization consists of the single feature row

$$\mathbf{v}_\alpha^\top = \psi(\overline{\mathbf{w}}_\alpha^\top \overline{\mathbf{X}}), \qquad \overline{\mathbf{w}}_\alpha = \begin{bmatrix} \alpha \\ -2\alpha \end{bmatrix}. \tag{413}$$

Since

$$\overline{\mathbf{w}}_\alpha^\top \overline{\mathbf{X}} = [\alpha \quad -2\alpha] \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} = [-3\alpha \quad -2\alpha \quad -\alpha], \tag{414}$$

the feature $\mathbf{v}_\alpha$ vanishes for every $\alpha > 0$. Hence,

$$\Omega_H(\boldsymbol{\theta}_\alpha) = \|\mathbf{H}\|_F^2 + \|\mathbf{A}\|_F^2 = \alpha^{-2} \longrightarrow 0 \qquad \text{as } \alpha \to \infty. \tag{415}$$

Consequently, Ω_H can be made arbitrarily small within this width-1 symmetry orbit. Since $\Omega_H \geq 0$, any minimum would therefore have to equal zero. But $\Omega_H = 0$ would require $\mathbf{A} = \mathbf{0}$ and hence force the realized function to vanish everywhere, contradicting the fact that $\psi(x-2)$ is nonzero for $x > 2$. Therefore, this symmetry orbit contains no MRNP.

We next relate this failure of existence to the positivity assumption $g_q > 0$ in Proposition F.8. The single neuron represents the unique essential parameter class q of the symmetry orbit. Choosing the unit-norm representative

$$\overline{\mathbf{w}}_q = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \tag{416}$$

the two features generated on $\mathbf{X}$ by the two orientations of q are

$$\mathbf{f}_q^+ = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{f}_q^- = \frac{1}{\sqrt{5}} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}. \tag{417}$$

Hence,

$$g_q^+ = 0, \qquad g_q^- > 0, \qquad g_q = 0, \tag{418}$$

so the positivity assumption in Proposition F.8 fails.

Moreover, the residual constraint uniquely fixes $t_q = 1$. This split also satisfies the orientation restriction in Equation (393), since $g_q^+ < g_q^-$, and requires only one neuron. Thus, the corresponding split is feasible at width 1, even though the representation-norm objective has no minimizer in the width-1 orbit. This shows that feasibility of $\mathcal{T}_{N_h}^H$ alone does not guarantee existence of an MRNP when $g_q = 0$.

The reason $g_q = 0$ can prevent existence of an MRNP is that it breaks the usual trade-off between feature magnitude $\|\mathbf{v}_j\|$ and readout magnitude $\|\mathbf{a}_j\|$ under positive scaling. For a neuron j in an essential class q, the representation-norm objective contains the two contributions $\|\mathbf{v}_j\|^2$ and $\|\mathbf{a}_j\|^2$. Ordinarily, positive scaling trades off these two terms: scaling up a neuron's incoming parameters $\overline{\mathbf{w}}_j$ increases $\|\mathbf{v}_j\|$, while simultaneously scaling down its readout weights $\mathbf{a}_j$ decreases $\|\mathbf{a}_j\|$, producing a finite optimum. When $g_q = 0$, however, an orientation of class q generates the zero feature on $\mathbf{X}$, so any neuron in that orientation satisfies $\|\mathbf{v}_j\| = 0$ regardless of the scale of its incoming parameters. The incoming parameters can therefore be scaled up arbitrarily while the readout weights $\mathbf{a}_j$ approach $\mathbf{0}$, driving the neuron's contribution to $\|\mathbf{A}\|_F^2$ toward zero without increasing its contribution to $\|\mathbf{H}\|_F^2$. The condition $g_q > 0$ rules out precisely this failure of the scaling trade-off.

G Numerical and analytical characterization of ReLU networks solving the XOR task

This appendix provides mathematical details supporting the ReLU networks illustrated in Figure 1. We first specify the XOR dataset, the two-neuron ReLU architecture, and the binary cross-entropy (BCE) training objective used throughout the analysis (Appendix G.1). We then describe how the six qualitatively distinct ReLU solutions to the XOR task shown in Panel B of Figure 1 were identified through a gradient-descent sweep (Appendix G.2), derive closed-form expressions for these solutions in a geometric parameterization (Appendix G.3), and prove that each can approach zero BCE loss in the limit of parameter rescaling (Appendix G.4).

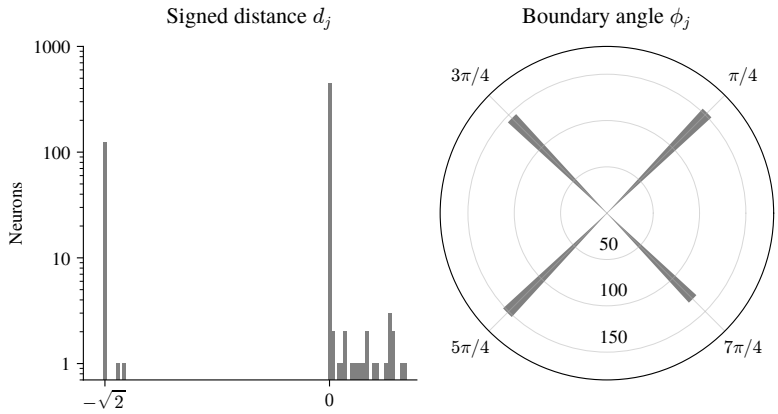

Figure G.1: **Boundary geometry of the** 592 **hidden neurons across** 296 **converged networks**. Ticks indicate the canonical values of the six solution types listed in Table G.1. Histogram counts are shown on a logarithmic scale for the signed distances d_j (left) and a linear scale for the angles ϕ_j (right). Of the 592 neurons, 19 have signed distances distinct from both dominant values, while 2 have angles distinct from all four dominant values.

G.1 Setup

The XOR dataset consists of the $P = 4$ points $\{-1, +1\}^2$, where same-sign inputs receive label 0 and opposite-sign inputs receive label 1:

$$\mathbf{X} = \begin{bmatrix} -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \end{bmatrix}, \qquad \mathbf{y} = [0 \quad 1 \quad 1 \quad 0]\,. \tag{419}$$

We consider two-neuron ReLU networks with scalar output,

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq a_1\,\psi(\mathbf{w}_1^\top \mathbf{x} + b_1) + a_2\,\psi(\mathbf{w}_2^\top \mathbf{x} + b_2) + b, \qquad \psi(z) = \max(0, z), \tag{420}$$

where $\mathbf{w}_j \in \mathbb{R}^2$, $b_j \in \mathbb{R}$ are the incoming weights and hidden biases, $a_j \in \mathbb{R}$ are the readout weights, $b \in \mathbb{R}$ is the output bias, and $\boldsymbol{\theta} = (\psi; \mathbf{w}_1, \mathbf{w}_2, b_1, b_2, a_1, a_2, b)$ denotes the network's parameterization. The output bias b lies outside the model class introduced in Section 2. It does not enter the hidden-activation matrix and remains unchanged under the parameter symmetries applied to these networks in Figures 1 and 2. The results of Sections 3 to 5 therefore apply directly to the hidden layer, whose contribution to the network output is $f_{\boldsymbol{\theta}} - b$. Training minimizes the mean BCE loss

$$\mathcal{L}(\boldsymbol{\theta}) \coloneqq -\frac{1}{4}\sum_{\mu=1}^{4}\big[y^\mu \log p^\mu + (1 - y^\mu)\log(1 - p^\mu)\big], \qquad p^\mu = \varsigma(f_{\boldsymbol{\theta}}(\mathbf{x}^\mu)), \tag{421}$$

where $\varsigma(z) \coloneqq (1 + \mathrm{e}^{-z})^{-1}$ denotes the logistic sigmoid.

Geometric parameterization. Each hidden neuron with nonzero incoming weights $\mathbf{w}_j \neq \mathbf{0}$ admits a geometric parameterization in terms of the angle ϕ_j and signed distance d_j of its activation boundary, together with a gain factor $g_j > 0$. Concretely, the pre-activation of the jth neuron is

$$z_j(\mathbf{x}) \coloneqq g_j\big(\mathbf{n}(\phi_j)^\top \mathbf{x} - d_j\big), \qquad \mathbf{n}(\phi) \coloneqq (\cos\phi, \sin\phi)^\top, \tag{422}$$

so that the zero level set $z_j(\mathbf{x}) = 0$ is the line with unit normal $\mathbf{n}(\phi_j)$ at signed distance d_j from the origin. This corresponds to the usual affine parameterization via $\mathbf{w}_j = g_j\,\mathbf{n}(\phi_j)$ and $b_j = -g_j d_j$.

G.2 Gradient-descent sweep and solution clustering

We trained `N_SEEDS` $= 1000$ two-neuron ReLU networks on the XOR dataset from independent random initializations, drawing each weight and bias i.i.d. from $\mathcal{U}(-1/\sqrt{2}, +1/\sqrt{2})$, corresponding to the default initialization in Equinox's `nn.Linear` module. Networks were optimized with full-batch Adam at learning rate $\eta = 0.1$ for `N_STEPS` $= 10^7$ steps, minimizing the BCE loss detailed in

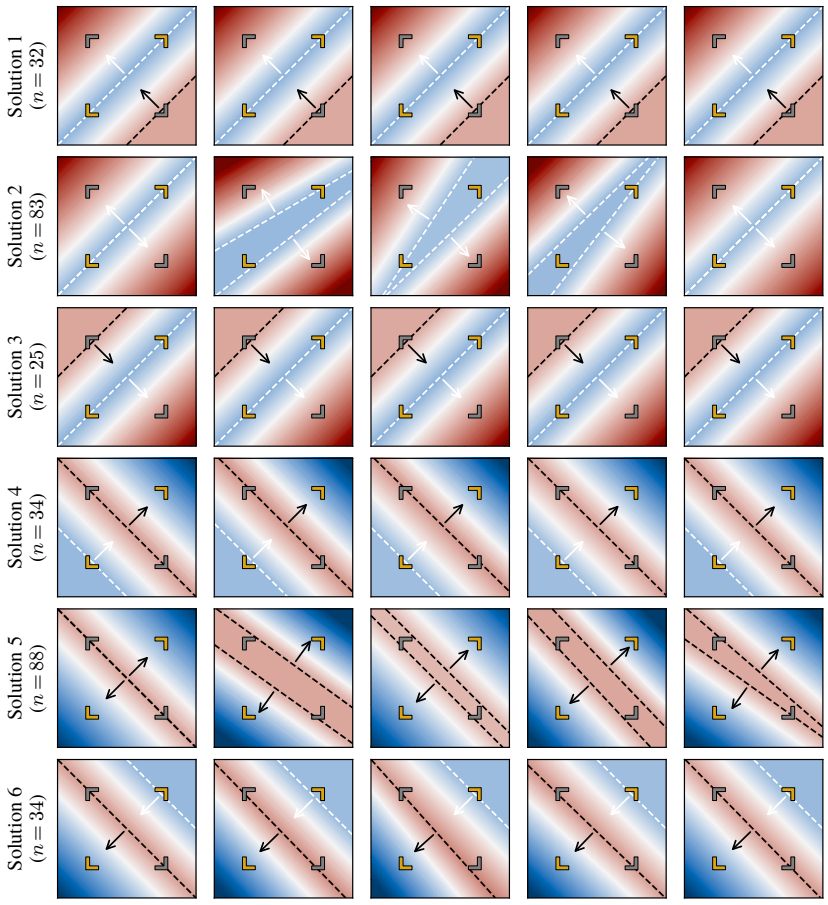

Figure G.2: **Cluster medoids and maximally dissimilar members**. The leftmost panel in each row shows the cluster medoid; moving right, each panel shows the member farthest from those already selected. Because training produces substantially larger readout weights than the closed-form representatives in Table G.1, the readouts of each network are rescaled for visualization to match the logit scale of Equation (426).

Equation (421). A run was deemed converged if its final loss fell below `TARGET_LOSS` $= 10^{-12}$; 296 of the 1000 runs converged. The sweep was run locally on the JAX CPU backend on a MacBook Air with an Apple M4 chip, 10 CPU cores, and 16 GB of unified memory; it required approximately 9 minutes of wall-clock time and used at most 0.5 GB of peak process memory. All experiments were implemented in JAX [v0.10.0, 54] using the Equinox neural-network library [v0.13.8, 55].

For each converged network, we converted the affine hidden parameters to the geometric parameterization of Equation (422):

$$g_j = \|\mathbf{w}_j\|, \qquad \phi_j = \operatorname{atan2}(w_{j,2}, w_{j,1}), \qquad d_j = -\frac{b_j}{g_j}. \tag{423}$$

Before comparing networks, we accounted for two structural symmetries: the positive-scaling symmetry of ReLU and the permutation symmetry of hidden neurons. We removed positive scaling from the neuron representation and quotiented out hidden-neuron permutations in the network-level distance. For the former, we absorbed the gain g_j into the readout and defined the effective readout weight $\beta_j := a_j g_j$. Each hidden neuron is then represented by

$$\mathbf{s}_j := (\cos\phi_j, \sin\phi_j, d_j, \beta_j), \tag{424}$$

where the unit-circle representation of ϕ_j avoids branch-cut artifacts at $\phi = \pm\pi$. A network is thus represented by its two neuron descriptors $\mathbf{s}_1, \mathbf{s}_2$ and its output bias b.

Table G.1: Closed-form representatives of the six ReLU solution types for XOR identified through the gradient-descent sweep. Each row specifies the angles (ϕ_1, ϕ_2), signed distances (d_1, d_2), readout weights (a_1, a_2), and output bias b. All solutions listed here use unit gain.

#	Family	(ϕ_1, ϕ_2)	(d_1, d_2)	(a_1, a_2)	b
1	Diagonal	$(\frac{3\pi}{4}, \frac{3\pi}{4})$	$(0, -\sqrt{2})$	$(2, -1)$	$\frac{1}{\sqrt{2}}$
2	Diagonal	$(\frac{3\pi}{4}, \frac{7\pi}{4})$	$(0, 0)$	$(1, 1)$	$-\frac{1}{\sqrt{2}}$
3	Diagonal	$(\frac{7\pi}{4}, \frac{7\pi}{4})$	$(-\sqrt{2}, 0)$	$(-1, 2)$	$\frac{1}{\sqrt{2}}$
4	Anti-diagonal	$(\frac{\pi}{4}, \frac{\pi}{4})$	$(-\sqrt{2}, 0)$	$(1, -2)$	$-\frac{1}{\sqrt{2}}$
5	Anti-diagonal	$(\frac{\pi}{4}, \frac{5\pi}{4})$	$(0, 0)$	$(-1, -1)$	$\frac{1}{\sqrt{2}}$
6	Anti-diagonal	$(\frac{5\pi}{4}, \frac{5\pi}{4})$	$(0, -\sqrt{2})$	$(-2, 1)$	$-\frac{1}{\sqrt{2}}$

For two networks $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, let $\mathbf{s}_j(\boldsymbol{\theta})$ and $\mathbf{s}_j(\boldsymbol{\xi})$ denote their neuron descriptors and let $b_{\boldsymbol{\theta}}$ and $b_{\boldsymbol{\xi}}$ denote their output biases. We define

$$d_{\mathrm{quot}}(\boldsymbol{\theta}, \boldsymbol{\xi}) := \min_{\pi \in \mathfrak{S}_2} \Big((b_{\boldsymbol{\theta}} - b_{\boldsymbol{\xi}})^2 + \sum_{j=1}^{2} \|\mathbf{s}_j(\boldsymbol{\theta}) - \mathbf{s}_{\pi(j)}(\boldsymbol{\xi})\|^2 \Big)^{1/2}. \tag{425}$$

This distance optimally matches the two hidden neurons before computing their Euclidean distance and therefore removes the hidden-neuron permutation symmetry. Because the neuron descriptors are themselves invariant under positive scaling, d_{quot} is invariant under both symmetries.

Before computing distances, we standardized each descriptor dimension to zero mean and unit variance to prevent differences in scale from dominating the comparison. In particular, the unit-circle coordinates are bounded in $[-1, 1]$, whereas signed distances and effective readout weights span wider ranges. For each neuron-level dimension, the standardization statistics were pooled across both hidden-neuron positions, ensuring that the same transformation is applied regardless of neuron ordering and preserving the permutation invariance of d_{quot}. We then performed average-linkage hierarchical agglomerative clustering using d_{quot} and cut the resulting dendrogram at the largest gap in its sequence of merge distances. This yielded six clusters, with 88, 83, 34, 34, 32, and 25 networks. Each cluster corresponds to one of the six solution types listed in Table G.1. Solutions 5 and 2, whose two activation boundaries pass through the origin, account for the two largest clusters, with 88 and 83 networks, followed by solutions 4, 6, 1, and 3 with 34, 34, 32, and 25 networks, respectively. Figure G.2 shows a small set of diverse representatives of each cluster, selected using medoid-seeded farthest-first (maxmin) sampling.

The corresponding six solution types form two families of three, which we call *diagonal* and *anti-diagonal* according to the orientation of their activation boundaries. Representatives are shown in Panel B of Figure 1, with all six solution types displayed individually in Figure G.3. These types characterize the dominant solutions obtained in this sweep; we do not claim that they exhaust the two-neuron ReLU XOR solutions attainable under other optimizers, learning rates, initialization distributions, or training protocols.

G.3 Closed-form characterization of the six dominant solution types

Each of the six solution types identified by gradient descent admits a convenient closed-form representative using the geometric parameterization introduced in Equation (422). We choose representatives with unit gain ($g_1 = g_2 = 1$), so that $\mathbf{w}_j = \mathbf{n}(\phi_j)$ and $b_j = -d_j$. Table G.1 lists their complete specifications, and Figure G.3 visualizes the networks they define.

Reduction to one dimension. A key structural property shared by all six solutions is that the two weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are collinear. Specifically:

- Solutions 1–3 have angles $\phi_j \in \{3\pi/4, 7\pi/4\}$. Since $\mathbf{n}(7\pi/4) = -\mathbf{n}(3\pi/4)$, both weight vectors lie on the line spanned by $\mathbf{n}(3\pi/4)$. The network output depends on $\mathbf{x}$ only through the *diagonal* projection $u_d = \mathbf{n}(3\pi/4)^\top \mathbf{x} = (x_2 - x_1)/\sqrt{2}$.

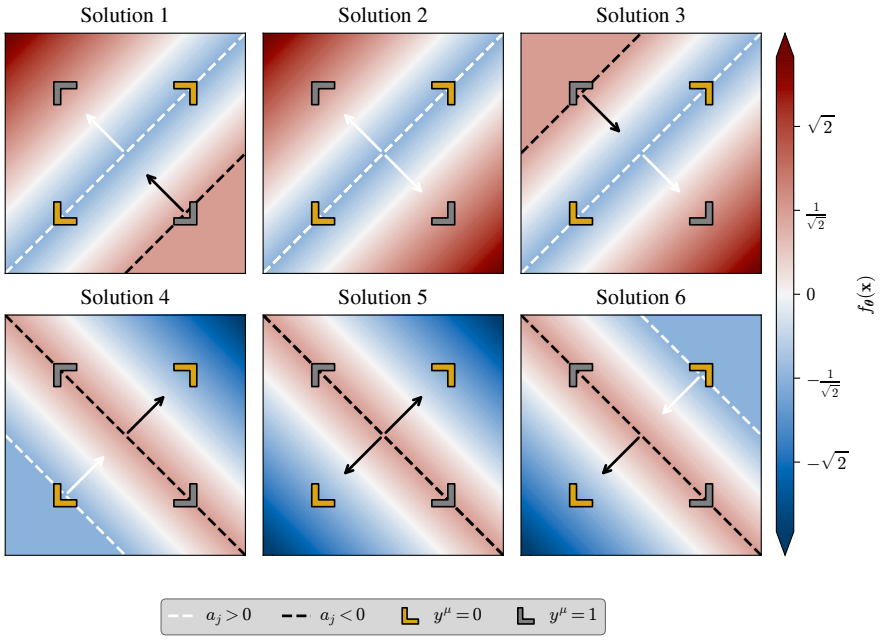

Figure G.3: **The six closed-form ReLU solutions of Table G.1 at unit gain**. Solutions 1–3 (top row) form the diagonal family, while solutions 4–6 (bottom row) form the anti-diagonal family. Within each family, the activation boundaries share a common orientation, while the solutions differ in their signed distances d_j and readout weights a_j. Each panel shows the logit $f_{\boldsymbol{\theta}}(\mathbf{x})$ of Equation (420) as a heatmap, the activation boundaries $z_j(\mathbf{x}) = 0$ as dashed lines with unit normals $\mathbf{n}(\phi_j)$, and the four XOR data points. All six networks assign every data point a logit of magnitude $1/\sqrt{2}$.

- Solutions 4–6 use angles $\phi_j \in \{\pi/4, 5\pi/4\}$, and the output depends only on the *anti-diagonal* projection $u_a = \mathbf{n}(\pi/4)^\top \mathbf{x} = (x_1 + x_2)/\sqrt{2}$.

In both cases, the four XOR points project to exactly three values $u \in \{-\sqrt{2}, 0, +\sqrt{2}\}$, with two points collapsing onto $u = 0$. For the diagonal family, the label-1 points project to $u_d = \pm\sqrt{2}$ and the label-0 points to $u_d = 0$. For the anti-diagonal family, the roles are reversed, with the label-0 points at $u_a = \pm\sqrt{2}$ and the label-1 points at $u_a = 0$.

Logits on the XOR data. One can verify by direct computation that every solution in Table G.1 produces the same logit values on the four XOR points:

$$f_{\boldsymbol{\theta}}(\mathbf{x}^\mu) = \begin{cases} +\frac{1}{\sqrt{2}}, & y^\mu = 1, \\ -\frac{1}{\sqrt{2}}, & y^\mu = 0, \end{cases} \qquad \mu = 1, \ldots, 4. \tag{426}$$

That is, all six networks classify XOR correctly, with every data point receiving the same logit magnitude $1/\sqrt{2}$. We illustrate this for one solution per family.

Solution 2 (diagonal family): The network computes

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \psi(u_d) + \psi(-u_d) - \frac{1}{\sqrt{2}} = |u_d| - \frac{1}{\sqrt{2}}, \tag{427}$$

where $u_d = (x_2 - x_1)/\sqrt{2}$. On the data: $f = -1/\sqrt{2}$ at $u_d = 0$ (label 0) and $f = +1/\sqrt{2}$ at $u_d = \pm\sqrt{2}$ (label 1).

Solution 5 (anti-diagonal family): The network computes

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = -\psi(u_a) - \psi(-u_a) + \frac{1}{\sqrt{2}} = -|u_a| + \frac{1}{\sqrt{2}}, \tag{428}$$

where $u_a = (x_1 + x_2)/\sqrt{2}$. On the data: $f = +1/\sqrt{2}$ at $u_a = 0$ (label 1) and $f = -1/\sqrt{2}$ at $u_a = \pm\sqrt{2}$ (label 0).

G.4 Approaching zero BCE loss by parameter scaling

Because the logistic sigmoid maps $\mathbb{R}$ into $(0, 1)$ without reaching the endpoints, no finite parameter vector $\boldsymbol{\theta}$ achieves exactly zero BCE loss. However, the parameter vector $\boldsymbol{\theta}$ of each of the six solutions in Table G.1 sits on a *scaling ray* along which the loss decreases monotonically to zero.

Proposition G.1. *Let $\boldsymbol{\theta}$ be any of the six solutions from Table G.1, and define the scaled parameterization*

$$\boldsymbol{\theta}_\tau := (\psi; \tau\mathbf{w}_1, \tau\mathbf{w}_2, \tau b_1, \tau b_2, a_1, a_2, \tau b), \qquad \tau > 0. \tag{429}$$

Then, $f_{\boldsymbol{\theta}_\tau}(\mathbf{x}^\mu) = \tau\, f_{\boldsymbol{\theta}}(\mathbf{x}^\mu)$, for $\mu = 1, \ldots, 4$, and the BCE loss along this ray satisfies

$$\mathcal{L}(\tau) := \mathcal{L}(\boldsymbol{\theta}_\tau) = \log(1 + \exp(-\tau/\sqrt{2})), \tag{430}$$

which is strictly decreasing in τ, with $\lim_{\tau\to\infty} \mathcal{L}(\tau) = 0$.

Proof. Under the scaling in Equation (429), the network output satisfies $f_{\boldsymbol{\theta}_\tau}(\mathbf{x}) = \tau\, f_{\boldsymbol{\theta}}(\mathbf{x})$, since ReLU is positively homogeneous of degree 1. By Equation (426), the logit at each data point thus has magnitude $\tau/\sqrt{2}$ with the appropriate sign. For a label-1 point, the BCE contribution is $-\log\varsigma(\tau/\sqrt{2})$. For a label-0 point, it is $-\log(1 - \varsigma(-\tau/\sqrt{2})) = -\log\varsigma(\tau/\sqrt{2})$, where the last step uses the identity $1 - \varsigma(z) = \varsigma(-z)$ of the logistic sigmoid. Since all four terms are identical, the mean BCE is $\mathcal{L}(\tau) = \log(1 + \exp(-\tau/\sqrt{2}))$. The map $\tau \mapsto -\tau/\sqrt{2}$ is strictly decreasing, while $\exp$ and $\log$ are strictly increasing; hence $\mathcal{L}(\tau)$ is strictly decreasing. The fact that $\lim_{\tau\to\infty} \mathcal{L}(\tau) = 0$ is evident. □

NeurIPS Paper Checklist

(i) **Claims**

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

Answer: [Yes]

Justification: The abstract and Section 1 accurately state the paper's four main contributions, and explicitly limit the scope to fully connected one-hidden-layer networks that implement the same function and whose parameterizations are related through the cataloged parameter symmetries.

Guidelines:

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

(ii) **Limitations**

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

Answer: [Yes]

Justification: The **Completeness** paragraph in Section 7 explicitly clarifies that we do not claim the cataloged functional parameter symmetries to exhaust the full fiber of a realized function, and that our analysis does not address cross-layer degeneracies or other sources of representational variation not induced by these symmetries.

Guidelines:

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

(iii) **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theorems, propositions, lemmas, corollaries, and definitions are numbered and cross-referenced, and the relevant assumptions are stated alongside each result in the main text. Complete proofs are provided in the appendix.

Guidelines:

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

(iv) **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper's main claims are theoretical. The only computational component is a gradient-descent sweep used to enumerate qualitatively distinct ReLU solutions to XOR (Figures 1 and 2), which serves as an illustrative example of a nonlinear network in which function-preserving parameter symmetries can be used to manipulate representational geometry. Appendices G.1 and G.2 describe the task setup and software stack, and specify all training details (dataset, initialization, optimizer, learning rate, training horizon, convergence criterion, and post-processing procedure), the number of random seeds and converged runs with the resulting cluster sizes, and the compute required (CPU backend, hardware, wall-clock time, and peak process memory). Moreover, Appendix G.3 provides closed-form characterizations for each of the six identified solutions, so that the reported solutions can be verified without rerunning the sweep. Code reproducing all figures and experimental results is available at github.com/mrvnthss/symmetries-representational-geometry.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(i) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(ii) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(iii) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(iv) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

(v) **Open access to data and code**

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

Answer: [Yes]

Justification: Code for generating all figures in the paper and performing the gradient-descent sweep to identify the ReLU XOR solutions is available at github.com/mrvnthss/symmetries-representational-geometry.

Guidelines:

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

(vi) **Experimental setting/details**

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

Answer: [Yes]

Justification: Appendices G.1 and G.2 specify the relevant training details, including the dataset, initialization, optimizer, learning rate, training horizon, convergence criterion, and post-processing procedure. No separate test split is used, since the computational component concerns fitting the full XOR dataset as an illustrative example.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

(vii) **Experiment statistical significance**

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

Answer: [Yes]

Justification: The gradient-descent sweep used to identify qualitatively distinct ReLU solutions to XOR is an illustrative enumeration rather than an inferential experiment requiring error bars. Appendix G.2 reports the number of random seeds, the convergence criterion, the number of converged runs, and the resulting cluster sizes, which provide the relevant aggregate information for this experiment.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

(viii) **Experiments compute resources**

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

Answer: [Yes]

Justification: Appendix G.2 specifies the software stack, CPU backend, compute hardware, wall-clock time, and peak process memory required for the ReLU gradient-descent sweep.

Guidelines:

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

(ix) **Code of ethics**

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

Answer: [Yes]

Justification: The work is theoretical and does not involve human subjects, personal data, or sensitive data. We have reviewed the NeurIPS Code of Ethics and believe the work conforms to it.

Guidelines:

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

(x) **Broader impacts**

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

Answer: [N/A]

Justification: The contributions are foundational and not tied to any particular application, dataset, or deployment setting. We therefore do not identify a direct path to specific positive or negative societal impacts that would warrant discussion beyond the paper checklist.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

(xi) **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release models or datasets. The released code reproduces the paper's figures and synthetic XOR experiment, and the analytical results pose no foreseeable immediate risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

(xii) **Licenses for existing assets**

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

Answer: [Yes]

Justification: The gradient-descent sweep used to identify the six ReLU solution types for XOR (Appendix G.2) was implemented using JAX [54] and Equinox [55], both released under the Apache License 2.0. Additional dependencies used for optimization and figure generation are specified in the public code repository at github.com/mrvnthss/symmetries-representational-geometry.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

(xiii) **New assets**

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

Answer: [Yes]

Justification: The released code is documented and provided at the URL listed in item (v). It is released under the MIT license.

Guidelines:

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

(xiv) **Crowdsourcing and research with human subjects**

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

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

(xv) **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

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

(xvi) **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology, scientific analysis, or experimental pipeline.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.