Parameter symmetries determine representational geometry in overparameterized nonlinear networks

Marvin Theiss [1 2 3] **Lukas Braun** [4] **Andrew M. Saxe** [5] **Erin Grant** [6 7]

Abstract

Representations are routinely used in machine learning, psychology, and neuroscience to probe the computations of biological and artificial systems. Yet it remains unclear to what extent computation constrains representation in artificial neural networks. One key obstacle is that these networks admit *parameter symmetries*: transformations of the parameters that preserve function exactly while reshaping representational geometry. Here we show that known parameter symmetries act on representations through just three primitives: *addition*, *duplication*, and *scaling*. This yields a closed-form descriptor of representational geometry as a sum of task-linked features and symmetry-induced noise. This decomposition further provides analytic bounds on representational similarity under parameter symmetries, revealing when functionally equivalent networks can become arbitrarily *dis*similar. Finally, we identify *privileged* representational geometries, which weight features by their computational importance and recover a stable link between representation and computation. Overall, our results delineate when representation 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, like weights or activities within a layer, reveals information about the end-to-end computation that a network performs (Zeiler and Fergus, 2014; Olah et al., 2020; Saphra and Wiegreffe, 2024; Mueller et al., 2026). In neuroscience and cognitive science, representational similarity analysis and related techniques for comparing internal representations within and across biological and artificial neural networks share this assumption that the geometry of neural activity reveals the computations it instantiates, and thus is a powerful tool for interpreting neural function in terms of neural activity (Haxby et al., 2014; Kriegeskorte et al., 2008; Yamins and DiCarlo, 2016).

However, artificial neural networks, like biological neural networks (Fakhar et al., 2024; Albantakis et al., 2024), are degenerate, as many distinct parameter configurations (Kunin et al., 2021; Entezari et al., 2022; Şimşek et al., 2021) and therefore distinct patterns of neural activity (Hermann and Lampinen, 2020; Flesch et al., 2022; Farrell et al., 2023; Chou et al., 2025; Lampinen et al., 2026) can support the same function. Further, in two-layer linear networks, there exists an analytical *double* dissociation of function and representation: networks can compute identical functions while exhibiting distinct representational geometries, and conversely can exhibit identical representational geometries while computing distinct functions (Braun et al., 2025). If such a double dissociation were general, it would call into question whether representations *per se* are ill-suited to study computation. What remains open is the extent to which function and representation are dissociated in *nonlinear* neural networks, where nonlinearities provide additional flexibility in representation for a given function and the space of realizable functions is more complex.

Here, we examine how function underdetermines representation as a consequence of *parameter symmetries*, transformations of the parameterization of a nonlinear neural network that leave its function unchanged (Hecht-Nielsen, 1990; Sussmann, 1992; Chen et al., 1993; Neyshabur et al., 2015; Zhao et al., 2026). Parameter symmetries can be classified by the actions on parameters that generate them (Şimşek et al., 2021; Martinelli et al., 2024). We refine this classification and derive the exact action of each symmetry on representation, demonstrating that all known parameter symmetries reduce to a composition of just three primi-

[1]Work partially done while a project research intern at the Gatsby Unit, UCL. [2]University of Tübingen [3]International Max Planck Research School for Intelligent Systems (IMPRS-IS) [4]Allen Institute for Neural Dynamics [5]Gatsby Unit & Sainsbury Wellcome Centre, UCL [6]University of Alberta [7]Amii. Correspondence to: Marvin Theiss <marvin.theiss@uni-tuebingen.de>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

tives: *addition*, *duplication*, and *scaling* of features. These primitives permit extensive variability in representational geometry, demonstrating that functional equivalence need not imply representational alignment.

Yet, the feature-transform view also reveals a path to identifiability: suitable implementation-level constraints remove symmetry-induced degrees of freedom and restore identifiable representational geometry, yielding a nonlinear analogue of the corresponding linear result (Braun et al., 2025). These implementation-level constraints single out representations that calibrate features according to their downstream computational importance, thus defining optimal representational geometry even in the degenerate model class of nonlinear neural networks. Our results provide a theoretical foundation for relating computation and representation in nonlinear neural networks via sufficient conditions under which computation constrains representation.

2. Preliminaries and setting

Nonlinear networks. We focus on a single layer of a nonlinear and fully-connected (feedforward) network of arbitrary depth, parameterized by incoming weights $\mathbf{W} \in \mathbb{R}^{N_h \times N_i}$, biases $\mathbf{b} \in \mathbb{R}^{N_h}$, and readout weights $\mathbf{A} \in \mathbb{R}^{N_o \times N_h}$. Letting $\mathbf{w}_j^\top$, b_j, and $\mathbf{a}_j$ denote the jth row, entry, and column of $\mathbf{W}$, $\mathbf{b}$, and $\mathbf{A}$, respectively, the hidden layer under consideration gives rise to the local input-output map

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum\nolimits_j \mathbf{a}_j \, \sigma(\mathbf{w}_j^\top \mathbf{x} + b_j), \tag{1}$$

with activation function $\sigma \colon \mathbb{R} \to \mathbb{R}$ and parameter vector $\boldsymbol{\theta} \coloneqq (\sigma; \mathbf{w}_j, b_j, \mathbf{a}_j)_{j=1}^{N_h}$.

Hidden activations. Collect a set of P inputs $\mathbf{x}^\mu \in \mathbb{R}^{N_i}$, where $\mu = 1, \ldots, P$, and $\mathbf{h}^\mu \in \mathbb{R}^{N_h}$, the hidden activation elicited by the μth input $\mathbf{x}^\mu$, into the matrices

$$\begin{aligned} \mathbf{X} &\coloneqq [\mathbf{x}^1, \ldots, \mathbf{x}^P] \in \mathbb{R}^{N_i \times P}, \\ \mathbf{H} &\coloneqq [\mathbf{h}^1, \ldots, \mathbf{h}^P] \in \mathbb{R}^{N_h \times P}. \end{aligned} \tag{2}$$

The matrix $\mathbf{H}$ of hidden activations is the central object in the study of representational geometry in artificial neural networks: its μth column $\mathbf{h}^\mu$ is the population response to the single input $\mathbf{x}^\mu$, while its jth row $\mathbf{z}_j^\top$ contains the hidden activity of the jth neuron across all inputs.

Representational geometry. One widely used descriptor of representational geometry is the uncentered Gram matrix $\mathbf{H}^\top \mathbf{H} \in \mathbb{R}^{P \times P}$ whose entries are dot products between the neural population responses elicited by pairs of inputs (Edelman, 1998; Kriegeskorte and Kievit, 2013). We refer to this matrix as the *representational similarity matrix (RSM)*. Since many measures of representational similarity depend on hidden activations only through $\mathbf{H}^\top \mathbf{H}$ (Kriegeskorte et al., 2008; Kornblith et al., 2019; Williams, 2024), the RSM is the natural object for studying variation in representational geometry. While the RSM is most widely framed as a matrix of pairwise similarities between stimulus-evoked activity patterns, we will instead view it as a sum of rank-one contributions from individual neurons, i.e.,

$$\begin{aligned} \mathrm{RSM} = \mathbf{H}^\top \mathbf{H} &= \sum_{j=1}^{N_h} \mathbf{z}_j \mathbf{z}_j^\top, \\ \mathbf{z}_j^\top &= \sigma(\mathbf{w}_j^\top \mathbf{X} + b_j \mathbf{1}^\top) \in \mathbb{R}^{1 \times P}. \end{aligned} \tag{3}$$

This decomposition makes clear that each hidden neuron contributes a rank-one term to the RSM. Our analysis proceeds by tracking how these contributions are reshaped under transformations of the layer's parameterization $\boldsymbol{\theta} = (\sigma; \mathbf{W}, \mathbf{b}, \mathbf{A})$ that leave the layer's local input-output map invariant.

3. Parameter symmetries in overparameterized neural networks

A layer map $f_{\boldsymbol{\theta}}$ may be realized by many different parameterizations $\boldsymbol{\theta}$ with distinct hidden activation matrices $\mathbf{H}$ and corresponding RSMs. Here, we study the resulting multiplicity in representational geometry through *parameter symmetries*. Such symmetries admit several definitions, distinguished by how strictly they preserve network behavior (Zhao et al., 2026). We focus throughout on the strictest, namely *functional parameter symmetries*: transformations of the layer's parameterization $\boldsymbol{\theta}$ that preserve the function $f_{\boldsymbol{\theta}}$ exactly, for every possible input. Any variation in representational geometry we identify therefore arises despite *exact* functional equivalence and cannot be attributed to input sampling.

3.1. A catalog of parameter symmetries

The parameter symmetries we consider take three qualitatively different forms. Some act on *individual* neurons, such as the positive scaling symmetry of positively homogeneous activations like ReLU (Neyshabur et al., 2015), which inversely scales a neuron's incoming and outgoing weights, or the sign-flip symmetry of odd activations like tanh (Hecht-Nielsen, 1990). Others exploit redundancy among *groups* of neurons that compute the same layer map through different configurations, such as *zero groups* whose readouts sum to zero or *duplicate groups* that distribute a single neuron's computation across multiple copies (Şimşek et al., 2021). A third class couples groups of neurons through algebraic structure of the activation σ, such as *linear groups* that can arise when σ decomposes into a sum of an even and a linear function, as is the case for ReLU and other commonly used activations (Martinelli et al., 2024). Panel A of Figure 1 illustrates all parameter symmetries we consider, building

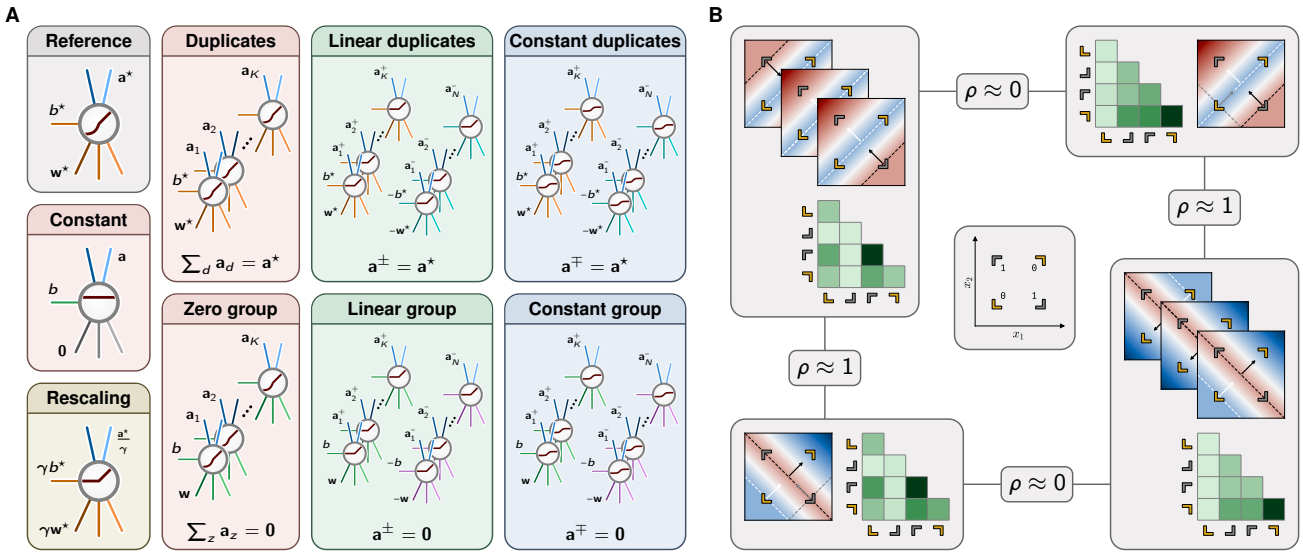

Figure 1. **Parameter symmetries can doubly dissociate function and representation**. **(A)** Parameter symmetries enable artificial neural networks to implement the same input-output function $f_{\boldsymbol{\theta}}$ through different parameterizations $\boldsymbol{\theta}$. The symmetries considered here differ qualitatively in how they act on a network's parameterization. Rescaling symmetries, such as the positive scaling symmetry of ReLU networks and the sign-flip symmetry of tanh networks, rescale the parameters of *individual* neurons. Other symmetries exploit redundancy among *groups* of neurons by distributing the computation of a single neuron across multiple copies (duplicates, zero groups). A further class of symmetries couples groups of neurons through algebraic structure in the activation function σ (linear duplicates, linear groups, constant duplicates, constant groups). **(B)** These parameter symmetries can doubly dissociate representation and function in nonlinear networks. Training one-hidden-layer ReLU networks with two neurons on XOR (center) using gradient descent from small initialization yields six distinct solutions, shown as heatmaps over input space with neuron decision boundaries overlaid. These solutions form two families (top left and bottom right) that solve the same task using different functions $f_{\boldsymbol{\theta}}$. Solutions from different families can have nearly uncorrelated RSMs, showing that networks solving the same task need not use similar representations. Moreover, overparameterized networks that compute the *same* function can have uncorrelated RSMs (top and bottom), whereas networks that implement *different* functions can exhibit perfectly correlated RSMs (left and right).

on and refining the taxonomies of Şimşek et al. (2021) and Martinelli et al. (2024). Full definitions are deferred to Appendix E, alongside an extensive classification of activation functions according to which symmetries they admit (Table F.1).

3.2. Irreducibility and orbits

The parameter symmetries described in Section 3.1 fall into two categories: *reparameterization* symmetries that preserve the width of the layer, and *overparameterization* symmetries that can change it by adding or removing neurons. We call a parameterization *irreducible* if no overparameterization symmetry can reduce its width without altering the function it computes, and *overparameterized* otherwise. A function $f_{\boldsymbol{\theta}^\star}$ can have many irreducible parameterizations $\boldsymbol{\theta}^\star$, since reparameterization symmetries leave width unchanged. The positive-scaling symmetry, for example, gives rise to continuous families of equivalent parameterizations. The collection of parameterizations reachable via parameter symmetries from an irreducible representative is precisely the set of functionally equivalent parameterizations across which we will analyze variations in representational geometry.

Definition 3.1 (Orbit of an irreducible representative)**.** Let $\boldsymbol{\theta}^\star$ be the parameterization of an irreducible hidden layer of width $N_h^\star$. The *orbit of* $\boldsymbol{\theta}^\star$ *at width* $N_h \geq N_h^\star$, denoted by $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$, is the set of all hidden-layer parameterizations of width N_h that can be obtained from $\boldsymbol{\theta}^\star$ through a finite, function-preserving composition of the parameter symmetries cataloged in Appendix E.

4. Parameter symmetries act on hidden features through three primitives

While two overparameterized hidden layers living in the same orbit generated by some irreducible representative $\boldsymbol{\theta}^\star$ compute the same input-output function, they may do so using drastically different hidden activations $\mathbf{H}$. As the degree of overparameterization $N_h/N_h^\star$ increases, the number of functionally equivalent parameterizations grows combinatorially, since additional hidden neurons introduce increasingly many ways of composing and distributing parameter symmetries while preserving the induced layer function. This raises the question of whether the resulting hidden activations $\mathbf{H}$ admit a tractable description, despite being generated by arbitrary compositions of parameter symme-

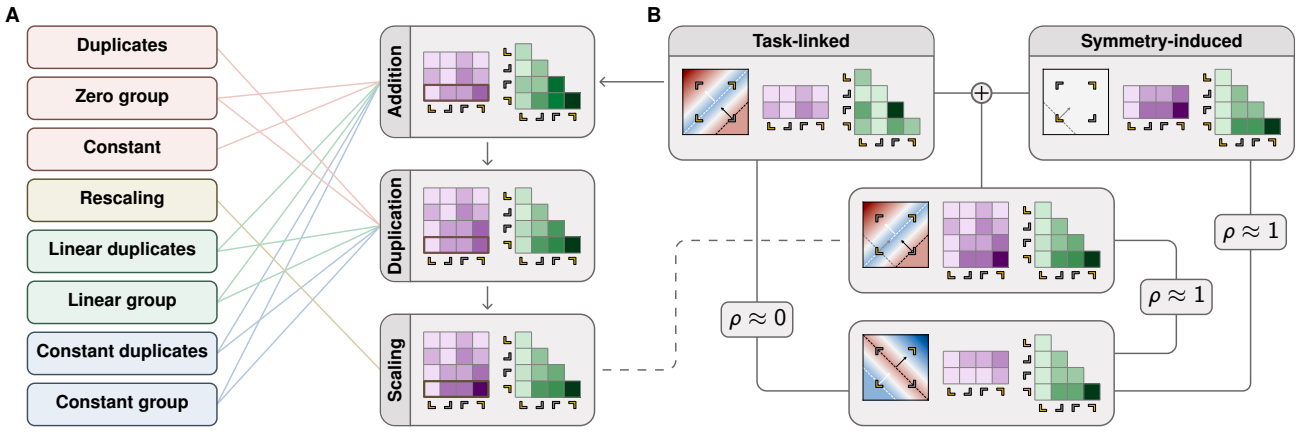

Figure 2. **Primitive transforms unify symmetries and isolate task-linked structure**. **(A)** Parameter symmetries and their decomposition into three primitive feature transforms: *addition*, *duplication*, and *scaling*. Shown are the effects on hidden activations and RSMs of introducing a zero group consisting of two neurons to one of the solutions in panel B of Figure 1, followed by rescaling one neuron, with the overall transformation decomposed into the three primitives. **(B)** Decomposing the full RSM of the resulting overparameterized network (center) into *task-linked* and *symmetry-induced* components reveals the source of the dissociation observed in Figure 1. Whereas the task-linked component is nearly uncorrelated with the RSM of a solution from the opposite family, the symmetry-induced component is nearly perfectly correlated with it, demonstrating the flexibility guaranteed by Proposition B.4.

tries. The key observation of this section is that they do: at the level of hidden activations, the apparent combinatorial complexity of the orbit reduces to a simple algebraic structure. In particular, we show that every hidden activation matrix induced by an element of the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ can be expressed in a canonical form involving only three primitive transformations applied to the hidden activations $\mathbf{H}^\star$ of the irreducible representative $\boldsymbol{\theta}^\star$.

4.1. From parameter symmetries to feature transforms

The parameter symmetries cataloged in Appendix E take rather different forms in parameter space: some introduce canceling groups of redundant neurons, some replace a reference neuron by a group of duplicates, and others rely on algebraic structure of the activation function to couple neurons with sign-flipped incoming parameters. We now show that, despite this diversity at the parameter level, the effects on hidden features induced by these symmetries reduce to just three primitive transformations.

Feature addition. Feature *addition* extends a given matrix of hidden activations $\mathbf{H}^\star$ by introducing the hidden features computed by additional neurons. Given a matrix $\mathbf{U} \in \mathbb{R}^{K\times P}$ representing these symmetry-induced features, the transformed hidden activation matrix can be expressed as

$$\mathbf{H}^{\mathbf{U}} = \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{U} \end{bmatrix}. \tag{4}$$

Feature duplication. Feature *duplication* replicates the hidden features of existing neurons. Given a duplication pattern $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{N_h^\star})^\top$, where $\nu_j \in \mathbb{N}_{>0}$ denotes the number of copies of the jth neuron present in the transformed layer, the transformed hidden activation matrix is given by

$$\mathbf{H}^{\boldsymbol{\nu}} = \mathbf{D}_{\boldsymbol{\nu}}\mathbf{H}^\star, \tag{5}$$

where $\mathbf{D}_{\boldsymbol{\nu}} \in \{0,1\}^{N_h \times N_h^\star}$ is a binary duplication matrix such that left-multiplication by $\mathbf{D}_{\boldsymbol{\nu}}$ repeats each row j of $\mathbf{H}^\star$ exactly ν_j times (see Appendix H.1.1 for its precise definition).

Feature scaling. Feature *scaling* rescales the hidden feature vector computed by each neuron by some factor. Given a vector of nonzero scaling factors $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{N_h^\star})^\top$, the rescaled hidden activation matrix can be written as

$$\mathbf{H}^{\boldsymbol{\alpha}} = \mathrm{diag}(\boldsymbol{\alpha})\mathbf{H}^\star. \tag{6}$$

Together, these primitives are sufficient both to translate the parameter-space symmetries considered here into feature-level transformations of hidden activations, and to characterize the set of hidden activation matrices $\mathbf{H}$ realizable by hidden layers in a fixed orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ in terms of the hidden activations $\mathbf{H}^\star$ of its irreducible representative $\boldsymbol{\theta}^\star$.

Proposition 4.1 (Feature-level characterization of symmetry orbits)**.** *Let $\boldsymbol{\theta}^\star$ be the parameterization of an irreducible hidden layer of width $N_h^\star$, and $\mathbf{H}^\star$ its hidden activation matrix. For any $N_h \geq N_h^\star$, every hidden activation matrix $\mathbf{H}$ induced by a layer $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ can be constructed from $\mathbf{H}^\star$ by a finite composition of feature* additions, duplications, *and* scalings. (Proof in Appendix H.1.2.)

4.2. A canonical form of the hidden activation matrix

Proposition 4.1 shows that every hidden activation matrix $\mathbf{H}$ in the orbit of $\boldsymbol{\theta}^\star$ can be obtained from $\mathbf{H}^\star$ by a finite composition of feature additions, duplications, and scalings. However, such compositions are not unique: the same $\mathbf{H}$ may arise from many different orderings and repetitions of the primitive transformations. The next result shows that this redundancy can be removed in the sense that any composition of primitives collapses to a canonical add–duplicate–scale expression.

Proposition 4.2 (Canonical form of hidden activations)**.** *Let $\boldsymbol{\theta}^\star$ be the parameterization of an irreducible hidden layer of width $N_h^\star$, and let $\mathbf{H}^\star \in \mathbb{R}^{N_h^\star \times P}$ denote its hidden activation matrix. For any $N_h \geq N_h^\star$, every hidden activation matrix $\mathbf{H}$ induced by a layer $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ can be written as*

$$\mathbf{H} = \mathrm{diag}(\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{U} \end{bmatrix}, \qquad \boldsymbol{\alpha} \in \mathbb{R}^{N_h}_{\neq 0}, \quad \boldsymbol{\nu} \in \mathbb{N}^{N_h^\star + K}_{>0}, \quad \mathbf{U} \in \mathbb{R}^{K \times P}, \tag{7}$$

where $\sum_{i=1}^{N_h^\star + K} \nu_i = N_h$ and $\mathbf{D}_{\boldsymbol{\nu}} \in \{0,1\}^{N_h \times (N_h^\star + K)}$ is the corresponding duplication matrix.

(Proof in Appendix H.1.4.)

This canonical form decomposes an arbitrary sequence of parameter symmetries applied to an irreducible hidden layer into three primitives. The matrix $\mathbf{U}$ contains all features introduced by overparameterization, $\mathbf{D}_{\boldsymbol{\nu}}$ records how irreducible and symmetry-induced features are duplicated, and $\mathrm{diag}(\boldsymbol{\alpha})$ captures feature-wise rescalings and sign flips.

5. Further results & discussion

It is common knowledge that function underdetermines the *parameterization* of artificial neural networks (Hecht-Nielsen, 1990; Sussmann, 1992; Kůrková and Kainen, 1994; Neyshabur et al., 2015; Dinh et al., 2017; Elbrächter et al., 2019; Phuong and Lampert, 2020). Building on work that makes this underdetermination precise via an enumeration of parameter symmetries in nonlinear networks (Şimşek et al., 2021; Martinelli et al., 2024), we examine instead the underdetermination in *representational geometry* (Section 4). We establish that representational geometry can vary considerably while function remains unchanged, supplying a theoretical bound on the underdetermination of representation by function (Section B) established empirically in prior work (Lampinen et al., 2024; Cloos et al., 2025; Bo et al., 2025; Lampinen et al., 2026). Last, we demonstrate that the coupling between representation and function can be restored for certain implementations characterized by efficiency constraints (Section C), mirroring a result from the linear regime (Braun et al., 2025). Though our work is theoretical at present, it holds consequences for the affordances of neural representations; we comment on three in Section D.

Code availability

All code used to produce the figures and experimental results presented in this paper has been made publicly available at: github.com/mrvnthss/symmetries-representational-geometry.

Acknowledgments

MT was supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). 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.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

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

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

Bansal, Yamini, Preetum Nakkiran, and Boaz 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.

Bo, Yiqing, Ansh Soni, Sudhanshu Srivastava, and Meenakshi 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.

Bradbury, James et al. (2018). *JAX: Composable transformations of Python+NumPy programs*. Version 0.10.0.

Braun, Lukas, Erin Grant, and Andrew 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.

Cai, Ming Bo, Nicolas W. Schuck, Jonathan W. Pillow, and Yael Niv (2019). "Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias". In: *PLOS Computational Biology* **15**(5), e1006299.

Cao, Rosa and Daniel L. K. Yamins (2024). "Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle". In: *Cognitive Systems Research* **85**, 101200.

Chen, An Mei, Haw-minn Lu, and Robert Hecht-Nielsen (1993). "On the geometry of feedforward neural network error surfaces". In: *Neural Computation* **5**(6), pp. 910–927.

Chen, Zirui and Michael F. Bonner (2025). "Universal dimensions of visual representation". In: *Science Advances* **11**(27), eadw7697.

Chou, Chi-Ning, Hang Le, Yichen Wang, and SueYeon 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.

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

Conwell, Colin 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.

Dinh, Lauren, Razvan Pascanu, Samy Bengio, and Yoshua 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.

Dujmović, Marin, Jeffrey S. Bowers, Federico Adolfi, and Gaurav Malhotra (2024). "Inferring DNN-brain alignment using representational similarity analyses can be problematic". In: *ICLR 2024 Workshop on Representational Alignment (Re-Align)*. Vienna, Austria.

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

Elbrächter, Dennis M., Julius Berner, and Philipp 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.

Entezari, Rahim, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur (2022). "The role of permutation invariance in linear mode connectivity of neural networks". In: *10th International Conference on Learning Representations*. Virtual Event.

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

Farrell, Matthew, Stefano Recanatesi, and Eric Shea-Brown (2023). "From lazy to rich to exclusive task representations in neural networks and neural codes". In: *Current Opinion in Neurobiology* **83**, 102780.

Feather, Jenelle, Meenakshi Khosla, Apurva R. Murty, and Aran Nayebi (Feb. 22, 2025). *Brain-model evaluations need the NeuroAI Turing test*. arXiv preprint.

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

Grigsby, Julia Elisenda, Kathryn Lindsey, and David 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.

Han, Yena, Tomaso A. Poggio, and Brian Cheung (2023). "System identification of neural systems: If we got it right, would we know?" In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. Honolulu, HI, USA: PMLR, pp. 12430–12444.

Haxby, James V., Andrew C. Connolly, and J. Swaroop Guntupalli (2014). "Decoding neural representational spaces using multivariate pattern analysis". In: *Annual Review of Neuroscience* **37**(1), pp. 435–456.

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

Hermann, Katherine and Andrew 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.

Holton, Eleanor et al. (2025). "Humans and neural networks show similar patterns of transfer and interference during

continual learning". In: *Nature Human Behaviour* **10**(1), pp. 111–125.

Huh, Minyoung, Brian Cheung, Tongzhou Wang, and Phillip 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.

Kidger, Patrick and Cristian 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.

Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey 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.

Kriegeskorte, Nikolaus and Rogier A. Kievit (2013). "Representational geometry: Integrating cognition, computation, and the brain". In: *Trends in Cognitive Sciences* **17**(8), pp. 401–412.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter A. Bandettini (2008). "Representational similarity analysis - connecting the branches of systems neuroscience". In: *Frontiers in Systems Neuroscience* **2**.

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

Kůrková, Věra and Paul C. Kainen (1994). "Functionally equivalent feedforward neural networks". In: *Neural Computation* **6**(3), pp. 543–558.

Lampinen, Andrew K., Stephanie C. Y. Chan, and Katherine Hermann (2024). "Learned feature representations are biased by complexity, learning order, position, and more". In: *Transactions on Machine Learning Research*.

Lampinen, Andrew Kyle, Stephanie C. Y. Chan, Yuxuan Li, and Katherine Hermann (2026). "Representation biases: Variance is not always a good proxy for importance". In: *eNeuro* **13**(3), ENEURO.0461-25.2026.

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

Linsley, Drew 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.

Martinelli, Flavio, Berfin Şimşek, Wulfram Gerstner, and Johanni 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.

Mueller, Aaron 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.

Neyshabur, Behnam, Russ R. Salakhutdinov, and Nati 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.

Olah, Chris et al. (2020). "Zoom in: An introduction to circuits". In: *Distill*.

Phuong, Mary and Christoph H. Lampert (2020). "Functional vs. parametric equivalence of ReLU networks". In: *8th International Conference on Learning Representations*. Virtual Event.

Rossem, Loek van and Andrew 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.

Saphra, Naomi and Sarah 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.

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

Şimşek, Berfin 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.

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

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

Tuckute, Greta, Jenelle Feather, Dana Boebinger, and Josh 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.

Williams, Alex H. (2024). "Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis". In: *Proceedings of UniReps: the Second Edition of the Workshop on Unify-*

ing Representations in Neural Models. Vol. 285. Proceedings of Machine Learning Research. Vancouver, Canada: PMLR, pp. 10–23.

Wolfram, Christopher and Aaron Schein (2025). "Layers at similar depths generate similar activations across LLM architectures". In: *2nd Conference on Language Modeling*. Montréal, Canada.

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

Zeiler, Matthew D. and Rob 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.

Zhao, Bo, Robin Walters, and Rose Yu (2026). "Symmetry in neural network parameter spaces". In: *Transactions on Machine Learning Research*.

Appendix Contents

A Notation **11**

B Parameter symmetries dissociate computation from representation **14**

B.1 Task-linked and symmetry-induced components of the RSM . . . 14

B.2 A cone-geometric characterization of representational geometries . . . 14

B.3 Symmetry-induced ambiguity in representational comparisons . . . 15

C Restoring identifiability via minimum-norm parameterizations **16**

C.1 Norm-minimizing constraints as implementation-level selection rules . . . 16

C.2 Minimum-norm implementations exhibit unique representational geometries . . . 16

C.3 Re-establishing the link between representation and computation . . . 17

C.4 When norm-minimization favors distributed computation . . . 17

D Discussion **17**

E Parameter symmetries in overparameterized nonlinear networks **18**

E.1 Generic reparameterization symmetries . . . 18

E.1.1 Permutation symmetry . . . 18

E.1.2 Positive scaling symmetry . . . 18

E.1.3 Sign-flip symmetry . . . 19

E.2 Activation-independent overparameterization symmetries . . . 19

E.2.1 Function-preserving realizations . . . 19

E.2.2 Minimality . . . 20

E.2.3 Distinctness . . . 20

E.3 Activation-dependent overparameterization symmetries . . . 20

E.3.1 Even-linear and constant-odd activations . . . 21

E.3.2 Aligned/opposite groups . . . 21

E.3.3 Overparameterization symmetries arising from even-linear activations . . . 21

E.3.4 Overparameterization symmetries arising from constant-odd activations . . . 22

E.3.5 Function-preserving realizations . . . 22

E.3.6 Choice of sign in aligned/opposite groups . . . 23

E.3.7 Nondegeneracy of activation-dependent symmetries . . . 23

F Symmetry properties of common activation functions **24**

F.1 Even-linear activations . . . 24

F.2 Constant-odd activations . . . 28

F.3 Positively homogeneous activations . . . 30

F.4 Activations with neither symmetry property 30

G Analytical XOR ReLU solutions **32**

G.1 Setup 32

G.2 Gradient-descent sweep 33

G.3 Analytical form of the six solutions 33

G.4 Approaching zero binary cross-entropy (BCE) loss via parameter scaling 34

H Proofs **35**

H.1 Proofs for Section 4 35

H.1.1 Duplication matrices 35

H.1.2 Primitive feature transforms generate hidden representations within orbits 37

H.1.3 Primitive feature transforms commute and collapse 38

H.1.4 Deriving a canonical form of the hidden activation matrix 40

H.2 Proofs for Section B 42

H.2.1 Decomposing RSMs into task-linked and symmetry-induced components 42

H.2.2 Characterizing representational geometries via convex cones 42

H.2.3 From cone freedom to ambiguity in similarity scores 43

H.3 Proofs for Section C 44

H.3.1 Setup 44

H.3.2 Row-wise decomposition of the norm-minimizing objectives 45

H.3.3 Weight-split minimization 46

H.3.4 One-dimensional minimization over the effective weight 47

H.3.5 Symmetry-induced hidden features and their RSM contributions 48

H.3.6 Identifiability of the RSM under MRNP and MWNP 49

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, γ_j, s_j) denote scalars,
- bold lowercase letters (e.g., $\mathbf{w}$, $\mathbf{a}$, $\mathbf{h}$) denote vectors,
- and bold uppercase letters (e.g., $\mathbf{W}$, $\mathbf{A}$, $\mathbf{H}$) denote matrices.

Quantities of the irreducible representative are distinguished from their (overparameterized) counterparts by a star and the use of the generic index j instead of i, e.g., $\mathbf{w}_j^\star$, $b_j^\star$, $\mathbf{a}_j^\star$ (irreducible) vs. $\mathbf{w}_i$, b_i, $\mathbf{a}_i$ (overparameterized).

Indices and dimensions

i, j	Generic neuron index
k	Index over symmetry-induced features
K	Number of symmetry-induced features
N_h	Width of the (overparameterized) hidden layer
$N_h^\star$	Width of the irreducible representative
N_i	Input dimension
N_o	Output dimension
P	Number of inputs in the analyzed input set
μ	Index over the P inputs

Inputs, hidden activations, and outputs

$\mathbf{h}^\mu \in \mathbb{R}^{N_h}$	Hidden representation elicited by input $\mathbf{x}^\mu$
$\mathbf{H} \in \mathbb{R}^{N_h \times P}$	Hidden activation matrix of the overparameterized layer
$\mathbf{H}^\star \in \mathbb{R}^{N_h^\star \times P}$	Hidden activation matrix of the irreducible representative
$\mathbf{u}_k^\top$	kth row of $\mathbf{U}$, symmetry-induced feature vector
$\mathbf{u}^+$, $\mathbf{u}^-$	Aligned and opposite induced feature rows generated by even-linear and constant-odd activation groups
$\mathbf{U} \in \mathbb{R}^{K \times P}$	Matrix collecting all K symmetry-induced features
$\mathbf{x}^\mu \in \mathbb{R}^{N_i}$	The μth input vector
$\mathbf{X} \in \mathbb{R}^{N_i \times P}$	Input matrix collecting all P inputs as columns
$\mathbf{z}_j^\top$	jth row of $\mathbf{H}^\star$, activity of irreducible neuron j across the input set

Network parameters

$\mathbf{a}_i$, $\mathbf{a}_j^\star$	Readout weights of single neuron
$\mathbf{a}^+$, $\mathbf{a}^-$	Aggregate readouts of aligned and opposite subgroups
$\mathbf{a}^\pm$	$\mathbf{a}^+ + \mathbf{a}^-$
$\mathbf{a}^\mp$	$\mathbf{a}^+ - \mathbf{a}^-$

$\mathbf{A}, \mathbf{A}^\star$	Readout-weight matrix
$\mathbf{b}, \mathbf{b}^\star$	Bias vector of the hidden layer
$b_i, b_j^\star$	Bias of single neuron
$f, f_{\boldsymbol{\theta}}$	Input-output map computed by the layer
$\mathbf{w}_i, \mathbf{w}_j^\star$	Incoming weights of single neuron
$\mathbf{W}, \mathbf{W}^\star$	Incoming-weight matrix of the hidden layer
$\boldsymbol{\theta}, \boldsymbol{\theta}^\star$	Parameter vector of the hidden layer

Activation functions

c	Constant value of the even component for a constant-odd activation
$e(x)$, $o(x)$	Even and odd components in the decomposition $\sigma(x) = e(x) + o(x)$
m	Slope of the linear odd component for an even-linear activation
$\sigma\colon \mathbb{R} \to \mathbb{R}$	Scalar activation function
$\sigma^{-1}(\{0\})$	Preimage of zero under σ

Parameter symmetry groups and orbits

$\mathcal{D}$	Index set of a duplicate-neuron group
$\mathcal{K} = \mathcal{N}^+ \sqcup \mathcal{N}^-$	Index set of an aligned/opposite group, partitioned into its aligned and opposite subgroups
$\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$	Orbit of $\boldsymbol{\theta}^\star$ at width N_h; the set of all parameterizations reachable from $\boldsymbol{\theta}^\star$ by function-preserving symmetries
$\mathfrak{S}_{N_h}$	Symmetric group acting on hidden neurons by permutation
$\mathcal{Z}$	Index set of a zero-neuron group
π	A permutation in $\mathfrak{S}_{N_h}$

Feature-transform primitives

$\mathcal{A}_{\mathbf{U}}, \mathcal{D}_{\boldsymbol{\nu}}, \mathcal{S}_{\boldsymbol{\alpha}}$	Feature-transform operators acting on the hidden activation matrix
$\mathbf{D}_{\boldsymbol{\nu}}$	Binary duplication matrix encoding $\boldsymbol{\nu}$
$\mathbf{H}^{\mathbf{U}}, \mathbf{H}^{\boldsymbol{\nu}}, \mathbf{H}^{\boldsymbol{\alpha}}$	Hidden activation matrices after addition, duplication, and scaling
$\boldsymbol{\alpha} \in \mathbb{R}_{\neq 0}^{N_h}$	Vector of nonzero per-neuron scaling factors
$\boldsymbol{\alpha}^2$	Hadamard square of $\boldsymbol{\alpha}$
$\boldsymbol{\gamma} = \mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2$	Effective per-feature weights of the overparameterized layer
γ_j	jth effective weight
$\gamma_j^\star$	Norm-minimizing effective weight for feature j
$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{N_h^\star})^\top$	Duplication pattern, $\nu_j \in \mathbb{N}_{>0}$ counts copies of neuron j

Representational similarity matrices

$\mathbf{M}_{[k]}$	k-truncated, $\mathcal{H}$-projected RSM
RSM	Representational similarity matrix $\mathbf{H}^\top\mathbf{H} \in \mathbb{R}^{P\times P}$
$\mathcal{S}_{[k]}(\mathbf{N})$	Set of attainable similarity scores against reference $\mathbf{N}$ over the cone $\mathcal{C}_{[k]}$
$\mathbf{u}_k\mathbf{u}_k^\top$	Rank-one contribution of symmetry-induced feature k to the RSM
$\mathbf{z}_j\mathbf{z}_j^\top$	Rank-one contribution of irreducible feature j to the RSM
$\rho(\mathbf{M},\mathbf{N})$	Pearson correlation between strict upper-triangular entries of two RSMs
$[\rho_-^{(k)},\rho_+^{(k)}]$	Closed interval bounding the attainable similarity scores at truncation level k

Cones and convex-analytic objects

$\mathrm{cone}(\cdot)$	Convex-conic hull operator
$\mathcal{C}_{[k]}$	Conic hull of the $\mathcal{H}$-projected rank-one terms
$\mathcal{H}$	Subspace of hollow symmetric matrices with zero off-diagonal mean
$\mathrm{Sym}(P)$	Space of $P\times P$ symmetric matrices
$\Pi_{\mathcal{H}}$	Orthogonal projection onto $\mathcal{H}$

Norms, inner products, and operators

$\lvert\cdot\rvert$	Absolute value, or set cardinality
$\arg\min$	Argument of the minimum
$\mathrm{diag}(\cdot)$	Diagonal matrix from a vector, or extraction of the diagonal of a matrix
$\langle\cdot,\cdot\rangle_F$	Frobenius inner product, $\langle\mathbf{M},\mathbf{N}\rangle_F = \mathrm{tr}(\mathbf{M}^\top\mathbf{N})$
$\lVert\cdot\rVert_F$	Frobenius norm
$\odot$	Hadamard product
$\lVert\cdot\rVert$	Euclidean norm
$\mathbf{0}$, $\mathbf{1}$, $\mathbf{I}$	Zero vector or matrix, all-ones vector, identity matrix
tr	Matrix trace
vech	Half-vectorization

Norm-minimization objectives

$\mathcal{J}_{\mathrm{MRNP}}(\boldsymbol{\theta})$	minimum representation-norm parameterization (MRNP) objective $\lVert\mathbf{H}\rVert_F^2 + \lVert\mathbf{A}\rVert_F^2$
$\mathcal{J}_{\mathrm{MWNP}}(\boldsymbol{\theta})$	minimum weight-norm parameterization (MWNP) objective $\lVert\mathbf{W}\rVert_F^2 + \lVert\mathbf{b}\rVert^2 + \lVert\mathbf{A}\rVert_F^2$
s_j	Feature-generating cost $\lVert\mathbf{z}_j\rVert^2$ (MRNP) or $\lVert\mathbf{w}_j^\star\rVert^2 + (b_j^\star)^2$ (MWNP)

B. Parameter symmetries dissociate computation from representation

Proposition 4.2 gives a canonical description of the hidden activations induced by the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$. We now use this description to analyze how function-preserving parameter symmetries affect representational geometry. First, we establish a decomposition of the RSM into task-linked and symmetry-induced components. We then recast this decomposition into cone-geometry, showing how feature addition enlarges the set of attainable representational geometries, while feature duplication and scaling select particular elements within that set. Finally, we translate these geometric degrees of freedom into ambiguity in representational comparisons, showing how a network with fixed function can attain a wide range of similarity scores independent of what it is compared against, an analytical statement consistent with prior work on failures in representational comparisons (Cai et al., 2019; Dujmović et al., 2024; Han et al., 2023; Bo et al., 2025; Lampinen et al., 2026).

B.1. Task-linked and symmetry-induced components of the RSM

We begin by deriving the RSM associated with the canonical hidden activation matrix of Proposition 4.2. The resulting expression separates the RSM into contributions from computationally relevant irreducible features and contributions from symmetry-induced added features.

Proposition B.1 (Canonical form of RSMs)**.** *Let $\boldsymbol{\theta}^\star$ be the parameterization of an irreducible hidden layer of width $N_h^\star$, and let $\mathbf{H}^\star \in \mathbb{R}^{N_h^\star \times P}$ denote its hidden activation matrix. Let $\mathbf{H}$ be the hidden activation matrix induced by a layer $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ in the canonical form of Equation (7). Denote the jth row of $\mathbf{H}^\star$ by $\mathbf{z}_j^\top$ and the kth row of $\mathbf{U}$ by $\mathbf{u}_k^\top$. Then the RSM of the overparameterized layer is*

$$\mathrm{RSM} = \mathbf{H}^\top \mathbf{H} = \sum_{j=1}^{N_h^\star} \gamma_j \, \mathbf{z}_j \mathbf{z}_j^\top + \sum_{k=1}^{K} \gamma_{N_h^\star + k} \, \mathbf{u}_k \mathbf{u}_k^\top, \qquad \boldsymbol{\gamma} = \mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2 \in \mathbb{R}_{>0}^{N_h^\star + K}, \tag{8}$$

where $\boldsymbol{\alpha}^2$ denotes the Hadamard square of $\boldsymbol{\alpha}$. (Proof in Appendix H.2.1.)

The canonical form of Proposition B.1 expresses every RSM in the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ as a weighted sum of rank-one terms, with weights $\boldsymbol{\gamma} = \mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2$ that the orbit leaves entirely unrestricted. On the irreducible features, feature scaling and duplication permit any positive reweighting of the rank-one terms $\mathbf{z}_j \mathbf{z}_j^\top$, with no constraint linking the weights to the role each feature plays in the layer's computation. Worse yet, feature addition admits further rank-one contributions $\mathbf{u}_i \mathbf{u}_i^\top$ from *symmetry-induced* features that are entirely redundant to the layer's computation.

B.2. A cone-geometric characterization of representational geometries

Proposition B.1 reveals that the same layer, computing the same function, gives rise to many different RSMs. We now capture the resulting freedom geometrically through a sequence of nested cones, indexed by the number k of symmetry-induced features. As is common practice (Williams, 2024), we discard self-similarities on the diagonal and mean-center off-diagonal entries, corresponding to the projection $\Pi_{\mathcal{H}}$ of RSMs onto the space $\mathcal{H}$ of hollow symmetric matrices defined formally in Section B.3.

Fix a parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ of width $N_h \geq N_h^\star + K$, with hidden activation matrix written as in Proposition 4.2, and let $\mathbf{u}_1, \ldots, \mathbf{u}_K$ denote its symmetry-induced features. We write

$$\mathbf{M}_{[k]} \coloneqq \sum_{j=1}^{N_h^\star} \gamma_j \, \Pi_{\mathcal{H}}(\mathbf{z}_j \mathbf{z}_j^\top) + \sum_{i=1}^{k} \gamma_{N_h^\star + i} \, \Pi_{\mathcal{H}}(\mathbf{u}_i \mathbf{u}_i^\top), \tag{9}$$

for the *k-truncated RSM* that retains only the first k symmetry-induced rank-one contributions. Analogously, we let

$$\mathcal{C}_{[k]} \coloneqq \mathrm{cone}\Big(\{\Pi_{\mathcal{H}}(\mathbf{z}_j \mathbf{z}_j^\top)\}_{j=1}^{N_h^\star} \cup \{\Pi_{\mathcal{H}}(\mathbf{u}_i \mathbf{u}_i^\top)\}_{i=1}^{k} \Big) \tag{10}$$

denote the conic hull generated by the rank-one terms retained in $\mathbf{M}_{[k]}$.

Proposition B.2 (Nested cones)**.** *The cones $\mathcal{C}_{[k]}$ form a nested sequence*

$$\mathcal{C}_{[0]} \subseteq \mathcal{C}_{[1]} \subseteq \cdots \subseteq \mathcal{C}_{[K]}, \tag{11}$$

with strict inclusion $\mathcal{C}_{[k+1]} \supsetneq \mathcal{C}_{[k]}$ if and only if $\Pi_{\mathcal{H}}(\mathbf{u}_{k+1} \mathbf{u}_{k+1}^\top) \notin \mathcal{C}_{[k]}$. (Proof in Appendix H.2.2.)

Proposition B.2 expresses two distinct mechanisms by which overparameterization decouples representational geometry from computation. At $k = 0$, the cone $\mathcal{C}_{[0]}$ already contains a continuum of RSMs that weight the irreducible features arbitrarily: the weights are free positive scalars, with no constraint tying them to any computationally meaningful quantity. For $k > 0$, the cone may grow further to admit rank-one contributions from features that are entirely irrelevant to the computation: when the new generator $\Pi_{\mathcal{H}}(\mathbf{u}_{k+1}\mathbf{u}_{k+1}^{\top})$ already lies in $\mathcal{C}_{[k]}$, the cone is unchanged but the RSM can still move within it; when it lies outside, the cone strictly enlarges, and the same layer computing the same function can exhibit representational structure entirely decoupled from the computation it performs.

B.3. Symmetry-induced ambiguity in representational comparisons

Representational similarity is often quantified by the Pearson correlation between the strict upper triangular entries of two RSMs. The cone freedom of Proposition B.2 translates directly into freedom in this similarity score: the same layer, computing the same function, can yield many different scores against the same reference geometry, and this arbitrariness worsens with overparameterization.

The following lemma justifies our use throughout this section of $\mathcal{H}$, the space of hollow symmetric matrices with zero mean off-diagonal entries (Appendix H.2.2): Pearson correlation between the strict upper triangular entries equals cosine similarity in $\mathcal{H}$.

Lemma B.3 (Pearson correlation as cosine similarity in $\mathcal{H}$)**.** *Let $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{P\times P}$ be symmetric matrices with $\Pi_{\mathcal{H}}(\mathbf{A}) \neq \mathbf{0}$ for $\mathbf{A} \in \{\mathbf{M}, \mathbf{N}\}$. The Pearson correlation between the strict upper triangular entries of $\mathbf{M}$ and $\mathbf{N}$ equals the cosine similarity of their projections onto $\mathcal{H}$:*

$$\rho(\mathbf{M}, \mathbf{N}) = \frac{\langle \Pi_{\mathcal{H}}(\mathbf{M}), \Pi_{\mathcal{H}}(\mathbf{N})\rangle_F}{\|\Pi_{\mathcal{H}}(\mathbf{M})\|_F \, \|\Pi_{\mathcal{H}}(\mathbf{N})\|_F}. \tag{12}$$

(Proof in Appendix H.2.3.)

For each k, define

$$\mathcal{S}_{[k]}(\mathbf{N}) \coloneqq \{\rho(\mathbf{M}, \mathbf{N}) \mid \mathbf{M} \in \mathcal{C}_{[k]},\ \mathbf{M} \neq \mathbf{0}\}, \tag{13}$$

the set of similarity scores attainable as $\mathbf{M}$ ranges over $\mathcal{C}_{[k]}$.

Proposition B.4 (Symmetry-induced ambiguity)**.** *Let σ be positively homogeneous of degree 1, and let $\mathbf{N} \in \mathbb{R}^{P\times P}$ be symmetric positive semidefinite with $\Pi_{\mathcal{H}}(\mathbf{N}) \neq \mathbf{0}$. Then for each k, the set $\mathcal{S}_{[k]}(\mathbf{N})$ is a closed interval, and these intervals are nested:*

$$[\rho_-^{(0)}, \rho_+^{(0)}] \subseteq [\rho_-^{(1)}, \rho_+^{(1)}] \subseteq \cdots \subseteq [\rho_-^{(K)}, \rho_+^{(K)}]. \tag{14}$$

(Proof in Appendix H.2.3.)

Proposition B.4 sharpens the dichotomy of Proposition B.2 into a statement about similarity scores. Feature addition that strictly enlarges $\mathcal{C}_{[k]}$ widens the interval, opening the range of achievable scores against any fixed reference; feature scaling and duplication move the score within a fixed interval. The same function thus admits a non-trivial range of representational similarity scores against any fixed reference, with the range potentially widening as overparameterization admits further symmetry-induced features.

The mechanism behind Proposition B.4 is clearest in the exclusive or (XOR) example illustrated in Figure 2. A single overparameterization step raises the cross-family similarity score from near zero to near one, without changing the function computed by the network. The irreducible solution (panel B, top left) contributes only one non-trivial rank-one generator, so $\mathcal{C}_{[0]}$ is a single ray and $\mathcal{S}_{[0]}$ is a singleton. Its similarity score to any cross-family solution is therefore fixed at $\rho \approx -0.03$; feature rescaling and duplication cannot change it. By contrast, adding a zero-neuron group whose feature reproduces the cross-family generator introduces a new generator outside $\mathcal{C}_{[0]}$, strictly enlarging the cone to $\mathcal{C}_{[1]}$ (Proposition B.2). Rescaling in this enlarged cone moves the RSM across rays in $\mathcal{C}_{[1]}$ and drives the cross-family similarity to $\rho \approx 0.99$. Panel B makes the dissociation explicit: the task-linked component (top left), which remains in $\mathcal{C}_{[0]}$, is nearly uncorrelated with the cross-family solution (bottom center), whereas the full overparameterized RSM (center), lying in $\mathcal{C}_{[1]}$, is nearly perfectly correlated with it. This is exactly the ambiguity guaranteed by Proposition B.4; explicit parameterizations are given in Appendix G.3.

C. Restoring identifiability via minimum-norm parameterizations

Section B paints a pessimistic picture: in overparameterized layers, representational geometry can vary substantially without any change in the underlying computation, and similarity to a fixed reference can range over increasingly large intervals. A similar dissociation between computation and representation is known to exist in two-layer linear networks (Braun et al., 2025). In the linear setting, suitable implementation-level selection rules restore representational identifiability. We now show that an analogous identifiability result holds for the nonlinear setting.

C.1. Norm-minimizing constraints as implementation-level selection rules

The ambiguity identified above arises because representational geometry is not determined by computation alone, but by the particular implementation realizing it. Restoring the link between computation and representation therefore requires restricting the class of admissible implementations. Following (Braun et al., 2025), we consider two norm-minimizing selection rules.

Definition C.1 (MRNP)**.** Fix an irreducible representative $\boldsymbol{\theta}^\star$. A parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ is a *minimum representation-norm parameterization (MRNP)* if it minimizes

$$\mathcal{J}_{\mathrm{MRNP}}(\boldsymbol{\theta}) \coloneqq \|\mathbf{H}\|_F^2 + \|\mathbf{A}\|_F^2 \tag{15}$$

over the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ at width $N_h \geq N_h^\star$.

Definition C.2 (MWNP)**.** Fix an irreducible representative $\boldsymbol{\theta}^\star$. A parameterization $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ is a *minimum weight-norm parameterization (MWNP)* if it minimizes

$$\mathcal{J}_{\mathrm{MWNP}}(\boldsymbol{\theta}) \coloneqq \|\mathbf{W}\|_F^2 + \|\mathbf{b}\|^2 + \|\mathbf{A}\|_F^2 \tag{16}$$

over the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ at width $N_h \geq N_h^\star$.

Both objectives select among functionally equivalent implementations by imposing a balance between feature generation and feature readout. The MRNP criterion enforces this balance in activation space by penalizing the hidden activation matrix together with the outgoing weights that read it out. The MWNP criterion enforces the analogous balance in parameter space by replacing the hidden-activation-matrix norm with the norms of the incoming weights and biases that generate the features. In both cases, the objective penalizes asymmetric implementations in which a hidden feature is suppressed at the hidden layer and compensated by a large readout weight, and vice versa.

C.2. Minimum-norm implementations exhibit unique representational geometries

We now show that, for hidden layers with positively homogeneous activations, the norm-minimizing selection rules of Section C.1 collapse the symmetry-induced family of RSMs within a fixed orbit to a single matrix. At the optimum, all symmetry-induced features are eliminated from the RSM, and each irreducible feature receives a unique effective weight with which it contributes to the computation.

Proposition C.3 (MRNP and MWNP select unique geometries)**.** *Let $\boldsymbol{\theta}^\star = (\sigma; \mathbf{w}_j^\star, b_j^\star, \mathbf{a}_j^\star)_{j=1}^{N_h^\star}$ be the parameterization of an irreducible hidden layer of width $N_h^\star$ with positively homogeneous activation function σ of degree 1, and let $\mathbf{H}^\star \in \mathbb{R}^{N_h^\star \times P}$ denote its hidden activation matrix. Let $\mathbf{z}_j^\top$ denote the jth row of $\mathbf{H}^\star$, and define*

$$s_j \coloneqq \begin{cases} \|\mathbf{z}_j\|^2, & \text{for MRNP} \\ \|\mathbf{w}_j^\star\|^2 + b_j^{\star 2}, & \text{for MWNP} \end{cases} \tag{17}$$

In the MRNP regime, assume additionally that $s_j > 0$ for all $j = 1, \ldots, N_h^\star$. Then every MRNP, respectively MWNP, in the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ at width $N_h \geq N_h^\star$ has the same RSM, namely

$$\mathrm{RSM} = \sum_{j=1}^{N_h^\star} \gamma_j^\star \, \mathbf{z}_j \mathbf{z}_j^\top, \qquad \gamma_j^\star = \operatorname*{arg\,min}_{\gamma \in (0,\infty)} \left(\gamma \, s_j + \gamma^{-1} \, \|\mathbf{a}_j^\star\|^2\right) = \frac{\|\mathbf{a}_j^\star\|}{\sqrt{s_j}}. \tag{18}$$

(Proof in Appendix H.3.6.)

Thus, both minimum-norm selection rules restore uniqueness at the level of representational geometry, but not necessarily at the level of parameterization.

C.3. Re-establishing the link between representation and computation

The optimal effective weights in Equation (18) solve a one-dimensional optimization problem: they each minimize the sum of a feature-generation cost, γs_j, and a readout cost, $\gamma^{-1}\|\mathbf{a}_j^\star\|^2$. Thus, the selected geometry weights each irreducible feature according to both how costly it is to represent and how strongly it is read out. For MRNP, where $s_j = \|\mathbf{z}_j\|^2$, the contribution of feature j to the selected RSM has magnitude

$$\|\gamma_j^\star \mathbf{z}_j \mathbf{z}_j^\top\|_F = \gamma_j^\star \|\mathbf{z}_j\|^2 = \frac{\|\mathbf{a}_j^\star\|}{\|\mathbf{z}_j\|}\|\mathbf{z}_j\|^2 = \|\mathbf{a}_j^\star\|\|\mathbf{z}_j\|, \qquad j = 1, \dots, N_h^\star. \tag{19}$$

By contrast, in the non-reweighted irreducible representative, the same feature contributes

$$\|\mathbf{z}_j \mathbf{z}_j^\top\|_F = \|\mathbf{z}_j\|^2, \qquad j = 1, \dots, N_h^\star, \tag{20}$$

which depends only on its activation norm and ignores the readout weights through which it affects the represented function. Under MRNP, the RSM contribution is instead proportional to the product of activation norm and readout norm: features with small activation norm but large readout norm are amplified, whereas features with large activation norm but small readout norm are attenuated. The MWNP objective admits an analogous interpretation with the activation norm $\|\mathbf{z}_j\|$ being replaced by the incoming parameter norm $(\|\mathbf{w}_j^\star\|^2 + b_j^{\star 2})^{1/2}$. In this sense, the minimum-norm objectives restore the link between computation and representation by replacing arbitrary symmetry-induced choices of scale or multiplicity with weights determined by each feature's contribution to the computation.

C.4. When norm-minimization favors distributed computation

Proposition C.3 selects a unique RSM per orbit but leaves open how optimal implementations distribute the work of representing each feature across hidden neurons. For positively homogeneous activations, this allocation problem is unconstrained: any duplication count ν_j with scaling factors satisfying $\sum_i \alpha_{j,i}^2 = \gamma_j^\star$ attains the optimum. For activations lacking positive homogeneity, increasing a feature's effective weight γ_j requires duplicating its generating neuron, with duplication splitting its readout contribution across copies and thereby reducing the squared readout-weight cost. Whenever $\|\mathbf{a}_j^\star\|/\sqrt{s_j} > \sqrt{2}$, this trade-off tips and the optimum attributes multiple neurons to the same feature (Appendix H.3). MRNP and MWNP thus reshape not only the representational geometry but also its implementation, preferring distributed encoding precisely when concentrating the representation on a single neuron would be expensive to read out.

D. Discussion

The preceding sections established that representational geometry can vary considerably while function remains unchanged (Section B), and that the coupling between representation and function can be restored for classes of implementations characterized by efficiency-related constraints (Section C). We now comment on the consequences of our work for neural representations, as announced in Section 5.

Completeness. Şimşek et al. (2021) establish a *complete* characterization of the global minima manifold for the teacher-student learning problem with restricted activation function and sufficient data, showing that *all* zero-loss solutions lie in the orbit of the irreducible teacher in the sense of Definition 3.1 but for a restricted set of symmetries. Martinelli et al. (2024) expand the set of symmetries to additional activation functions, but do not prove the orbit generated by these symmetries to be complete. Further, both papers focus on identifiability *within* a layer, and thus do not address degeneracies *across* layers, as arise for example from collapse of the representational geometry in an intermediate layer of a deep network (Grigsby et al., 2023, mechanism (iv)). As a consequence, we do not completely characterize all possible representations for a given function, and so provide an upper bound on representational similarity rather than a tight characterization.

Contravariance and task complexity. *Universal* representations are observed across artificial neural networks trained for different tasks and across different modalities, and even networks built with different architectures (Rossem and Saxe, 2024; Huh et al., 2024; Chen and Bonner, 2025). One possible explanation for universality is that the complexity of a function is *contravariant* to the "dispersion" (variability) of its implementations; that is, harder problems admit fewer solutions (Cao and Yamins, 2024). Contravariance would explain why artificial neural networks trained on an increasing range of tasks converge to universal representations (Li et al., 2016; Bansal et al., 2021; Wolfram and Schein, 2025). Yet, our results demonstrate that contravariance need not hold if the implementation is overparameterized and lacks additional implementation-level

constraints. In particular, in a proportional regime where the number of parameters of a model increases along with the complexity of the task on which it is trained (as arguably is the case in practice (Bahri et al., 2024)), the solution space of implementations need not narrow, contrary to contravariance. Contravariance, as a principle at the level of *function* rather than *implementation*, cannot alone account for universality.

Individual differences. Idiosyncrasies in representational geometry do persist, even among models of the same architecture (Schrimpf et al., 2020; Conwell et al., 2024; Tuckute et al., 2023; Linsley et al., 2023). Nevertheless, it is a goal of cognitive computational neuroscience to account for only the differences in neural activity that meaningfully separate individuals (Thobani et al., 2025; Feather et al., 2025), though the notion of meaningfulness should, of course, be contextual (Baker et al., 2026). Our characterization of the space of possible representational geometries in Section 4 distinguishes inter-individual differences that reflect arbitrary degeneracies from those that have computational consequences. In particular, the normative selection rules in Section C can separate individuals according to computational properties such as noise-robustness and transfer performance (Braun et al., 2025; Holton et al., 2025).

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

E.1. Generic reparameterization symmetries

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

E.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 an arbitrary permutation. Reordering the parameters of the layer according to

$$(\mathbf{w}_i, b_i, \mathbf{a}_i)_{i=1}^{N_h} \mapsto (\mathbf{w}_{\pi(i)}, b_{\pi(i)}, \mathbf{a}_{\pi(i)})_{i=1}^{N_h} \tag{21}$$

does not change the function computed by the layer, since

$$\sum_{i=1}^{N_h} \mathbf{a}_{\pi(i)}\, \sigma(\mathbf{w}_{\pi(i)}^\top \mathbf{x}_n + b_{\pi(i)}) = \sum_{i=1}^{N_h} \mathbf{a}_i\, \sigma(\mathbf{w}_i^\top \mathbf{x}_n + b_i). \tag{22}$$

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

E.1.2. Positive scaling symmetry

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

$$\sigma(\alpha x) = \alpha\, \sigma(x), \qquad \alpha > 0. \tag{23}$$

Rescaling the parameters of a single neuron via

$$(\mathbf{w}_i, b_i, \mathbf{a}_i) \mapsto (\alpha \mathbf{w}_i, \alpha b_i, \alpha^{-1} \mathbf{a}_i), \qquad \alpha > 0 \tag{24}$$

does not change the function computed by that neuron since

$$\alpha^{-1} \mathbf{a}_i\, \sigma(\alpha \mathbf{w}_i^\top \mathbf{x}_n + \alpha b_i) = (\alpha^{-1}\alpha) \mathbf{a}_i\, \sigma(\mathbf{w}_i^\top \mathbf{x}_n + b_i) = \mathbf{a}_i\, \sigma(\mathbf{w}_i^\top \mathbf{x}_n + b_i). \tag{25}$$

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

E.1.3. SIGN-FLIP SYMMETRY

Suppose the activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$ is odd, i.e.,

$$\sigma(-x) = -\sigma(x). \tag{26}$$

Flipping the sign of both the incoming parameters $(\mathbf{w}_i, b_i)$ and the readout weights $\mathbf{a}_i$, i.e.,

$$(\mathbf{w}_i, b_i, \mathbf{a}_i) \mapsto (-\mathbf{w}_i, -b_i, -\mathbf{a}_i) \tag{27}$$

does not change the function computed by that neuron since

$$-\mathbf{a}_i\, \sigma(-\mathbf{w}_i^\top \mathbf{x}_n - b_i) = -\mathbf{a}_i\, \sigma(-(\mathbf{w}_i^\top \mathbf{x}_n + b_i)) = \mathbf{a}_i\, \sigma(\mathbf{w}_i^\top \mathbf{x}_n + b_i). \tag{28}$$

We refer to this as *sign-flip symmetry*.

E.2. Activation-independent overparameterization symmetries

In contrast to the generic function-preserving reparameterization symmetries discussed in Appendix E.1, the symmetries considered in the next two 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. (2021) 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 E.2.1), and demonstrate that the symmetry classes are minimal (Appendix E.2.2) and mutually distinct (Appendix E.2.3).

Compared with prior formulations (Şimşek et al., 2021; Martinelli et al., 2024), we impose the following nondegeneracy condition to exclude decomposable cases.

Definition E.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 $\mathcal{J} \subsetneq \mathcal{I}$,

$$\sum_{j\in\mathcal{J}} \mathbf{a}_j \neq \mathbf{0}. \tag{29}$$

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

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

$$\sum_{z\in\mathcal{Z}} \mathbf{a}_z = \mathbf{0} \qquad \text{and} \qquad \{\mathbf{a}_z\}_{z\in\mathcal{Z}} \text{ is subset-nonzero.} \tag{30}$$

Definition E.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 2 hidden neurons with readout weights $\mathbf{a}_d$ satisfying $\mathbf{w}_d = \mathbf{w}^\star$ and $b_d = b^\star$ for all $d \in \mathcal{D}$. We call $\mathcal{D}$ a *duplicate-neuron group* if

$$\sum_{d\in\mathcal{D}} \mathbf{a}_d = \mathbf{a}^\star \qquad \text{and} \qquad \{\mathbf{a}_d\}_{d\in\mathcal{D}} \text{ is subset-nonzero.} \tag{31}$$

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

E.2.1. FUNCTION-PRESERVING REALIZATIONS

Zero-neuron groups. Adding a zero-neuron group to a hidden layer leaves the layer function unchanged. Indeed,

$$\sum_{z\in\mathcal{Z}} \mathbf{a}_z\, \sigma(\mathbf{w}_z^\top \mathbf{x} + b_z) = \sigma(\mathbf{w}^\top \mathbf{x} + b) \underbrace{\sum_{z\in\mathcal{Z}} \mathbf{a}_z}_{=\mathbf{0}} = \mathbf{0}. \tag{32}$$

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

$$\sum_{d\in\mathcal{D}} \mathbf{a}_d\, \sigma(\mathbf{w}_d^\top \mathbf{x} + b_d) = \sigma({\mathbf{w}^\star}^\top \mathbf{x} + b^\star) \sum_{d\in\mathcal{D}} \mathbf{a}_d = \mathbf{a}^\star\, \sigma({\mathbf{w}^\star}^\top \mathbf{x} + b^\star). \tag{33}$$

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

$$\mathbf{a}\, \sigma(\mathbf{w}^\top \mathbf{x} + b) = \mathbf{a}\, \sigma(b). \tag{34}$$

Thus, for the layer function to remain unchanged, the constant contributions of all such neurons must cancel. If σ is constant-odd (Definition E.6), this offset may also be compensated by constant-neuron groups (Definition E.10) or constant-duplicate-neuron groups (Definition E.11).

E.2.2. Minimality

Requiring that the shared incoming weight vector be nonzero for zero-neuron and duplicate-neuron groups, and incorporating the subset-nonzero condition from Definition E.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 $\mathcal{Z}$ be 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 E.2 singles out the smallest groups of neurons that could be removed from a hidden layer without altering its input-output mapping.

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* $\mathbf{a}^\star$,[1] 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

$$\sum_{d\in\mathcal{D}\setminus\mathcal{J}} \mathbf{a}_d = \sum_{d\in\mathcal{D}} \mathbf{a}_d - \sum_{j\in\mathcal{J}} \mathbf{a}_j = \mathbf{a}^\star - \mathbf{a}^\star = \mathbf{0}, \tag{35}$$

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.

E.2.3. Distinctness

Definitions E.2 to E.4 not only satisfy the minimality property discussed in the preceding subsection, 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 readout weights $\mathbf{a}^\star \neq \mathbf{0}$ to be nonzero. Hence, Definitions E.2 to E.4 are all mutually exclusive.

E.3. Activation-dependent overparameterization symmetries

The three activation-*independent* symmetries in Appendix E.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 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 E.3.1), and then formalize the notion of a group of neurons splitting into aligned and opposite subgroups (Appendix E.3.2). We then review the corresponding symmetry groups for even-linear activations

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

(Appendix E.3.3) and constant-odd activations (Appendix E.3.4), before discussing their function-preserving realizations (Appendix E.3.5), the choice of sign in aligned/opposite subgroups (Appendix E.3.6), and the degenerate cases excluded by the nondegeneracy requirements placed on aligned/opposite subgroups (Appendix E.3.7).

E.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 $\sigma(x) = e(x) + o(x)$, where

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2}, \qquad o(x) = \frac{\sigma(x) - \sigma(-x)}{2}, \tag{36}$$

satisfying $e(-x) = e(x)$ and $o(-x) = -o(x)$. Following (Martinelli et al., 2024), we distinguish two classes of activations by the structure of their even and odd components.

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

Definition E.6 (Constant-odd activations)**.** An activation function is called *constant-odd* if its even component is constant: $\sigma(x) = c + o(x)$ 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, $\sigma(x) = mx + c$, which includes linear networks.

E.3.2. Aligned/opposite groups

Here, we formalize subgroups of hidden neurons that share incoming weights and biases up to a sign flip as *aligned/opposite subgroups*. These underlie the activation-dependent symmetry groups of (Martinelli et al., 2024), 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 collapse into disjoint activation-*independent* symmetry groups. A detailed discussion of these degenerate cases is provided in Appendix E.3.7.

Definition E.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 index set $\mathcal{K}$ 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

$$(\mathbf{w}_k, b_k) = (\mathbf{w}, b), \quad k \in \mathcal{N}^+ \qquad \text{and} \qquad (\mathbf{w}_k, b_k) = (-\mathbf{w}, -b), \quad k \in \mathcal{N}^-. \tag{37}$$

An aligned/opposite group is called *nondegenerate* if it satisfies

1. $\{\mathbf{a}_k\}_{k\in\mathcal{N}^+}$ and $\{\mathbf{a}_k\}_{k\in\mathcal{N}^-}$ are subset-nonzero
2. $\mathbf{a}^{\pm} \notin \{\mathbf{a}^{\mp}, -\mathbf{a}^{\mp}\}$

for the aggregate weights

$$\mathbf{a}^+ := \sum_{k\in\mathcal{N}^+} \mathbf{a}_k, \qquad \mathbf{a}^- := \sum_{k\in\mathcal{N}^-} \mathbf{a}_k, \qquad \mathbf{a}^{\pm} := \mathbf{a}^+ + \mathbf{a}^-, \qquad \mathbf{a}^{\mp} := \mathbf{a}^+ - \mathbf{a}^-. \tag{38}$$

Note that the splitting is defined only up to a global sign flip: the same set of neurons splits into aligned subgroup $\mathcal{N}^-$ and opposite subgroup $\mathcal{N}^+$ with respect to parameters $(-\mathbf{w}, -b)$. In Appendix E.3.6 we show that for some of the symmetry groups introduced in Appendices E.3.3 and E.3.4 the labels "aligned" and "opposite" are indeed interchangeable, whereas for others they acquire a semantic meaning that removes this sign-ambiguity altogether.

E.3.3. Overparameterization symmetries arising from even-linear activations

Assume $\sigma(x) = e(x) + mx$ is even-linear in the sense of Definition E.5. For an aligned/opposite group with parameters $(\mathbf{w}, b)$, letting $z_k := \mathbf{w}_k^\top \mathbf{x} + b_k$ and $z := \mathbf{w}^\top \mathbf{x} + b$, 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

$$\sum_k \mathbf{a}_k\, \sigma(z_k) = \sum_k \mathbf{a}_k\, e(z_k) + \sum_k \mathbf{a}_k\, m z_k = \mathbf{a}^{\pm} e(z) + \mathbf{a}^{\mp} m z. \tag{39}$$

Hence, the aggregate readout weights $\mathbf{a}^{\pm}$ and $\mathbf{a}^{\mp}$ control the even and linear components of the group's contribution independently. In particular, an aligned/opposite group with $\mathbf{a}^{\pm} = \mathbf{0}$ generates a purely linear contribution, while an aligned/opposite group satisfying $\mathbf{a}^{\pm} = \mathbf{a}^{\star}$ replicates the activation of a single reference neuron with parameters $(\mathbf{w}, b, \mathbf{a}^{\star})$ up to a misaligned linear contribution. This motivates the following two symmetry groups, which correspond to the "even + linear" symmetries of (Martinelli et al., 2024).

Definition E.8 (Linear-neuron group)**.** Suppose $\sigma(z) = e(z) + mz$ is even-linear. A *linear-neuron group* is a nondegenerate aligned/opposite group with parameters $(\mathbf{w}, b)$ such that $\mathbf{a}^{\pm} = \mathbf{0}$.

Definition E.9 (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 nondegenerate aligned/opposite group with parameters $(\mathbf{w}^{\star}, b^{\star})$ such that $\mathbf{a}^{\pm} = \mathbf{a}^{\star}$.

E.3.4. Overparameterization symmetries arising from constant-odd activations

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

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \sum_k \mathbf{a}_k \, c + \sum_k \mathbf{a}_k \, o(z_k) = \mathbf{a}^{\pm} c + \mathbf{a}^{\mp} o(z). \tag{40}$$

Analogously to the even-linear case, $\mathbf{a}^{\pm}$ and $\mathbf{a}^{\mp}$ now control the constant and odd components independently, and the (misaligned) contribution of such an aligned/opposite group is constant for appropriate choices of $\mathbf{a}^{\mp}$. In particular, choosing $\mathbf{a}^{\mp} = \mathbf{0}$ generates a purely constant contribution, while an aligned/opposite group satisfying $\mathbf{a}^{\mp} = \mathbf{a}^{\star}$ replicates the activation of a single reference neuron with parameters $(\mathbf{w}, b, \mathbf{a}^{\star})$ up to a misaligned constant contribution. This motivates the following two symmetry groups, corresponding to the "odd (+ constant)" symmetries of (Martinelli et al., 2024).

Definition E.10 (Constant-neuron group)**.** Suppose $\sigma(z) = c + o(z)$ is constant-odd. A *constant-neuron group* is a nondegenerate aligned/opposite group with parameters $(\mathbf{w}, b)$ such that $\mathbf{a}^{\mp} = \mathbf{0}$.

Definition E.11 (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 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 E.4: the former involve aligned/opposite groups with nonzero incoming weights whereas the latter are individual neurons with vanishing incoming weights.

E.3.5. Function-preserving realizations

Even-linear activations. For even-linear $\sigma(x) = e(x) + mx$ (Definition E.5), the combined contribution of an aligned/opposite group with parameters $(\mathbf{w}, b)$ equals $\mathbf{a}^{\pm} e(z) + \mathbf{a}^{\mp} mz$, where $z := \mathbf{w}^{\top}\mathbf{x} + b$ (cf. Appendix E.3.3). Adding a linear-neuron group ($\mathbf{a}^{\pm} = \mathbf{0}$, Definition E.8) thus contributes a purely linear function of the input:

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \mathbf{a}^{\pm} e(z) + \mathbf{a}^{\mp} mz = \mathbf{a}^{\mp} m \, (\mathbf{w}^{\top}\mathbf{x} + b). \tag{41}$$

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 E.9) reproduces the reference neuron's contribution up to a residual linear term:

$$\sum_k \mathbf{a}_k \, \sigma(z_k) = \mathbf{a}^{\pm} e(z^{\star}) + \mathbf{a}^{\mp} mz^{\star} = \mathbf{a}^{\star} \, \sigma(z^{\star}) + (\mathbf{a}^{\mp} - \mathbf{a}^{\star}) \, mz^{\star}, \tag{42}$$

where $z^{\star} := \mathbf{w}^{\star\top}\mathbf{x} + b^{\star}$. For the layer function to remain unchanged, these linear contributions must cancel collectively across linear- and linear-duplicate-neuron groups, in direct analogy to the constant offsets of individual constant neurons (Definition E.4) in the activation-independent case.

Constant-odd activations. For constant-odd $\sigma(x) = c + o(x)$ (Definition E.6), 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 (cf. Appendix E.3.4). Adding a constant-neuron group ($\mathbf{a}^{\mp} = \mathbf{0}$, Definition E.10) thus contributes an input-independent offset:

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

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 E.11) reproduces the reference neuron's contribution up to a residual constant term:

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

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

E.3.6. Choice of sign in aligned/opposite groups

As noted after Definition E.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 non-duplicate 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.

Duplicate-neuron groups 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 non-duplicate groups), is semantically present but algebraically silent (for linear-duplicate groups), or is resolved by the defining equations themselves (for constant-duplicate groups).

E.3.7. Nondegeneracy of activation-dependent symmetries

The structural and nondegeneracy conditions integrated into Definition E.7, together with the requirement $\mathbf{a}^\star \neq \mathbf{0}$ in the duplicate variants, jointly ensure that the four activation-dependent symmetry classes (Definitions E.8 to E.11) form a minimal taxonomy that is mutually distinct and disjoint from the activation-independent classes of Appendix E.2. Building on the definitions of (Martinelli et al., 2024), we impose additional conditions that prevent activation-dependent groups from collapsing into one or more activation-independent groups, blurring into a different activation-dependent class, or decomposing into smaller groups of the same 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 E.4. The combined contribution to the layer reduces to $\mathbf{a}^\pm e(b) + \mathbf{a}^\mp m b$ 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, say $\mathcal{N}^- = \emptyset$, all neurons would share the parameters $(\mathbf{w}, b)$, and the four activation-dependent defining conditions would reduce to activation-independent ones: $\mathbf{a}^\pm = \mathbf{0}$ (linear-neuron) and $\mathbf{a}^\mp = \mathbf{0}$ (constant-neuron) both become $\sum_k \mathbf{a}_k = \mathbf{0}$, the defining condition of a zero-neuron group, while $\mathbf{a}^\pm = \mathbf{a}^\star$ (linear-duplicate) and $\mathbf{a}^\mp = \mathbf{a}^\star$ (constant-duplicate) both become $\sum_k \mathbf{a}_k = \mathbf{a}^\star$, the defining condition of a duplicate-neuron group. Requiring both subgroups to be nonempty thus prevents activation-dependent symmetries from coinciding with activation-independent ones.

Subset-nonzero readouts within each subgroup. Suppose the readouts $\{\mathbf{a}_k\}_{k \in \mathcal{N}^+}$ admit a nonempty proper subset $\mathcal{J} \subsetneq \mathcal{N}^+$ with $\sum_{j \in \mathcal{J}} \mathbf{a}_j = \mathbf{0}$. The neurons in $\mathcal{J}$ then form a zero-neuron group at $(\mathbf{w}, b)$, and removing them from $\mathcal{N}^+$ leaves $\mathbf{a}^+$ and $\mathbf{a}^-$, hence $\mathbf{a}^\pm$ and $\mathbf{a}^\mp$, unchanged. The smaller aligned/opposite group obtained in this way still satisfies the same defining condition, so the original group decomposes into a strictly smaller group of the same class together with a

zero-neuron subgroup, in violation of minimality. The same argument applies to $\mathcal{N}^-$. The subset-nonzero condition rules out such decompositions, in direct analogy to its role for activation-independent symmetries (Appendix E.2.2).

Mismatched aggregate readouts ($\mathbf{a}^\pm \notin \{\mathbf{a}^\mp, -\mathbf{a}^\mp\}$). The two cases $\mathbf{a}^\pm = \mathbf{a}^\mp$ and $\mathbf{a}^\pm = -\mathbf{a}^\mp$ correspond, respectively, to $\mathbf{a}^- = \mathbf{0}$ and $\mathbf{a}^+ = \mathbf{0}$, turning one of the two subgroups into a zero-readout block at $(\pm\mathbf{w}, \pm b)$ on its own. The activation-dependent symmetry group then splits into disjoint activation-independent pieces; for instance, a linear-duplicate-neuron group with $\mathbf{a}^- = \mathbf{0}$ decomposes into a duplicate-neuron group at $(\mathbf{w}^\star, b^\star)$ together with a zero-neuron group at $(-\mathbf{w}^\star, -b^\star)$, and the analogous decompositions hold for the remaining three classes. Excluding $\mathbf{a}^\pm \in \{\mathbf{a}^\mp, -\mathbf{a}^\mp\}$ ensures that both subgroups carry nontrivial readout mass and that the activation-dependent structure is irreducible. As a special case, this condition rules out groups with $\mathbf{a}^\pm = \mathbf{a}^\mp = \mathbf{0}$, which would otherwise satisfy both the linear-neuron and constant-neuron defining conditions simultaneously while contributing nothing to the layer.

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 non-duplicate counterparts. Moreover, a reference neuron with $\mathbf{a}^\star = \mathbf{0}$ contributes nothing to the layer output, so "duplicating" it would carry no semantic content. Requiring $\mathbf{a}^\star \neq \mathbf{0}$ keeps the duplicate and non-duplicate variants disjoint and ensures that duplicate variants genuinely replicate a nontrivial reference neuron.

F. Symmetry properties of common activation functions

Which of the overparameterization symmetries introduced in Appendix E can arise in a given hidden layer depends on algebraic properties of its activation function. Here, we classify every scalar activation available from the `jax.nn` module (v0.10.0, Bradbury et al., 2018), covering the standard activations commonly used in practice, according to whether it is even-linear (Appendix F.1), constant-odd (Appendix F.2), positively homogeneous of degree 1 (Appendix F.3), or none of the above (Appendix F.4). These classes are not mutually exclusive, and the complete classification is summarized in Table F.1.

F.1. Even-linear activations

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

GELU. For $\sigma(x) = x\,\Phi(x)$, where $\Phi(x) = \frac{1}{2}(1 + \mathrm{erf}(x/\sqrt{2}))$ is the standard normal CDF, the symmetry $\Phi(-x) = 1 - \Phi(x)$ gives

$$\sigma(-x) = -x\,\Phi(-x) = -x(1 - \Phi(x)). \tag{45}$$

The even component is not constant,

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{x\,\Phi(x) - x(1 - \Phi(x))}{2} = x\,\Phi(x) - \frac{x}{2}, \tag{46}$$

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

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = \frac{x\,\Phi(x) + x(1 - \Phi(x))}{2} = \frac{x}{2}, \tag{47}$$

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

Hard SiLU. By definition, $\mathrm{hard\_silu}(x) = x\ \mathrm{hard\_sigmoid}(x)$. Since the hard sigmoid satisfies $\mathrm{hard\_sigmoid}(x) = \frac{1}{2} + o(x)$ for some odd function o, we have

$$\mathrm{hard\_silu}(x) = x\,o(x) + \frac{x}{2}, \tag{48}$$

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

Table F.1. Symmetries of all scalar activation functions available from the `jax.nn` module (v0.10.0). The parameters α (`celu`, `elu`, `leaky_relu`, `selu`), λ (`selu`), and b (`squareplus`) are assumed to satisfy $\alpha > 0$, $\lambda \geq 1$, and $b > 0$. For `leaky_relu`, we assume $\alpha \neq 1$; when $\alpha = 1$, the activation reduces to the identity. The **Even+Lin** column gives the slope m of the linear component for even-linear activations; column **Const+Odd** provides the constant c for constant-odd activations. Odd functions, which are the ones exhibiting the sign-flip symmetry, are those with vanishing constant $c = 0$. **Scaling** refers to the positive scaling symmetry of positively homogeneous activations.

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

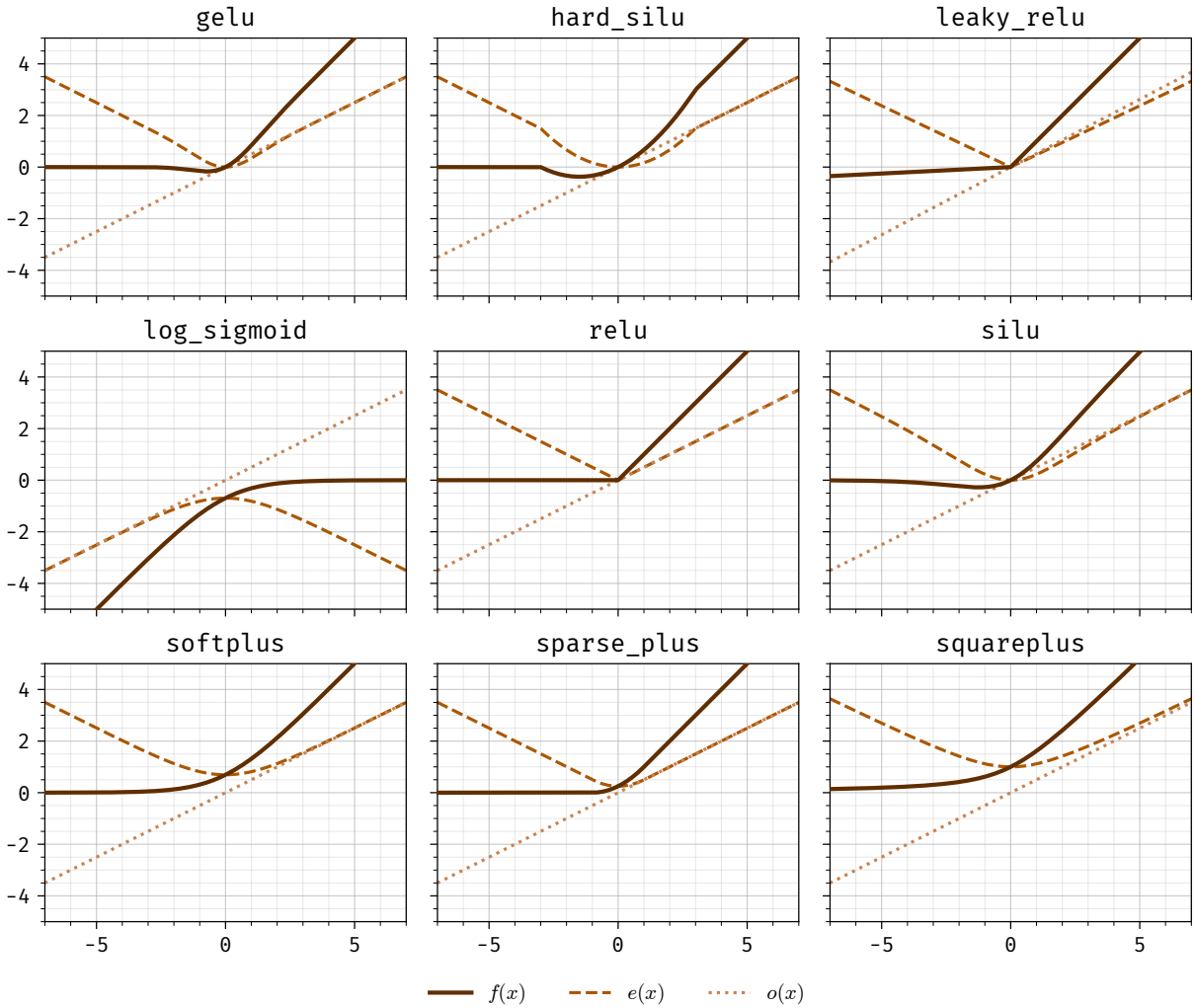

Figure F.1. Activations from `jax.nn` classified as even-linear, shown with their decomposition into even and odd components. Solid dark lines show the activations, dashed lines their even components, and dotted lines their odd components. In each case, the odd component is a line passing through the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

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

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

$$\sigma(x) = \begin{cases} \alpha x, & x < 0 \\ x, & x \geq 0 \end{cases} \quad \text{and} \quad \sigma(-x) = \begin{cases} -x, & x < 0 \\ -\alpha x, & x \geq 0 \end{cases} \tag{49}$$

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

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{x - \alpha x}{2} = \frac{1 - \alpha}{2} x, \tag{50}$$

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

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = \frac{x + \alpha x}{2} = \frac{1 + \alpha}{2} x, \tag{51}$$

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

$$\begin{aligned} e(x) &= \frac{\sigma(x)+\sigma(-x)}{2} = \frac{-\log(1+\mathrm{e}^{-x}) - \log(1+\mathrm{e}^{x})}{2} \\ &= -\frac{\log((1+\mathrm{e}^{-x})(1+\mathrm{e}^{x}))}{2} = -\frac{\log(2(1+\cosh(x)))}{2}, \end{aligned} \tag{52}$$

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

$$\begin{aligned} o(x) &= \frac{\sigma(x)-\sigma(-x)}{2} = \frac{-\log(1+\mathrm{e}^{-x}) + \log(1+\mathrm{e}^{x})}{2} \\ &= \frac{1}{2}\log\left(\frac{1+\mathrm{e}^{x}}{1+\mathrm{e}^{-x}}\right) = \frac{1}{2}\log\left(\frac{\mathrm{e}^{x}(\mathrm{e}^{-x}+1)}{1+\mathrm{e}^{-x}}\right) = \frac{\log(\mathrm{e}^{x})}{2} = \frac{x}{2}. \end{aligned} \tag{53}$$

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

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

$$e(x) = \frac{\sigma(x)+\sigma(-x)}{2} = \frac{1}{2}\begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases} = \frac{|x|}{2}, \tag{54}$$

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

$$o(x) = \frac{\sigma(x)-\sigma(-x)}{2} = \frac{x}{2}, \tag{55}$$

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

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

$$\mathrm{silu}(x) = x\,o(x) + \frac{x}{2}, \tag{56}$$

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

$$x\,o(x) = x\ \mathrm{sigmoid}(x) - \frac{x}{2} = \frac{x}{1+\mathrm{e}^{-x}} - \frac{x}{2}, \tag{57}$$

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

Softplus. Using the identity

$$\sigma(-x) = \log(1+\mathrm{e}^{-x}) = \log(1+\mathrm{e}^{x}) - x = \sigma(x) - x \tag{58}$$

of the softplus activation σ, we get

$$e(x) = \frac{\sigma(x)+\sigma(-x)}{2} = \sigma(x) - \frac{x}{2}, \tag{59}$$

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

$$o(x) = \frac{\sigma(x)-\sigma(-x)}{2} = \frac{x}{2}, \tag{60}$$

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

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

$$\sigma(x) = \begin{cases} 0, & x \le -1 \\ \frac{1}{4}(x+1)^2, & -1 < x < 1 \\ x, & x \ge 1 \end{cases} \quad \text{and} \quad \sigma(-x) = \begin{cases} -x, & x \le -1 \\ \frac{1}{4}(-x+1)^2, & -1 < x < 1 \\ 0, & x \ge 1 \end{cases} \tag{61}$$

For $|x| < 1$, we have

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{(x+1)^2 + (-x+1)^2}{8} = \frac{x^2+1}{4} \tag{62}$$

and

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = \frac{(x+1)^2 - (-x+1)^2}{8} = \frac{x}{2}. \tag{63}$$

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

Squareplus. The squareplus activation function satisfies

$$\sigma(-x) = \frac{-x + \sqrt{x^2+b}}{2}, \tag{64}$$

so the even component equals

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{(x + \sqrt{x^2+b}) + (-x + \sqrt{x^2+b})}{4} = \frac{\sqrt{x^2+b}}{2}, \tag{65}$$

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

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = \frac{(x + \sqrt{x^2+b}) - (-x + \sqrt{x^2+b})}{4} = \frac{x}{2}, \tag{66}$$

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

F.2. Constant-odd activations

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

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

$$\sigma(x) = \begin{cases} 0, & x \le -3 \\ (x+3)/6, & -3 < x < 3 \\ 1, & x \ge 3 \end{cases} \quad \text{and} \quad \sigma(-x) = \begin{cases} 1, & x \le -3 \\ (-x+3)/6, & -3 < x < 3 \\ 0, & x \ge 3 \end{cases} \tag{67}$$

From this, it follows that

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{(x+3) + (-x+3)}{12} = \frac{1}{2}, \qquad |x| < 3. \tag{68}$$

By inspection, the same is true for $|x| \ge 3$. Thus, the hard sigmoid activation function is constant-odd with $c = \frac{1}{2}$. The odd component satisfies $o(x) = -\frac{1}{2}$ for $x \le -3$. 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 $\sigma(x) = x$ is odd, so it is trivially constant-odd with constant $c = 0$.

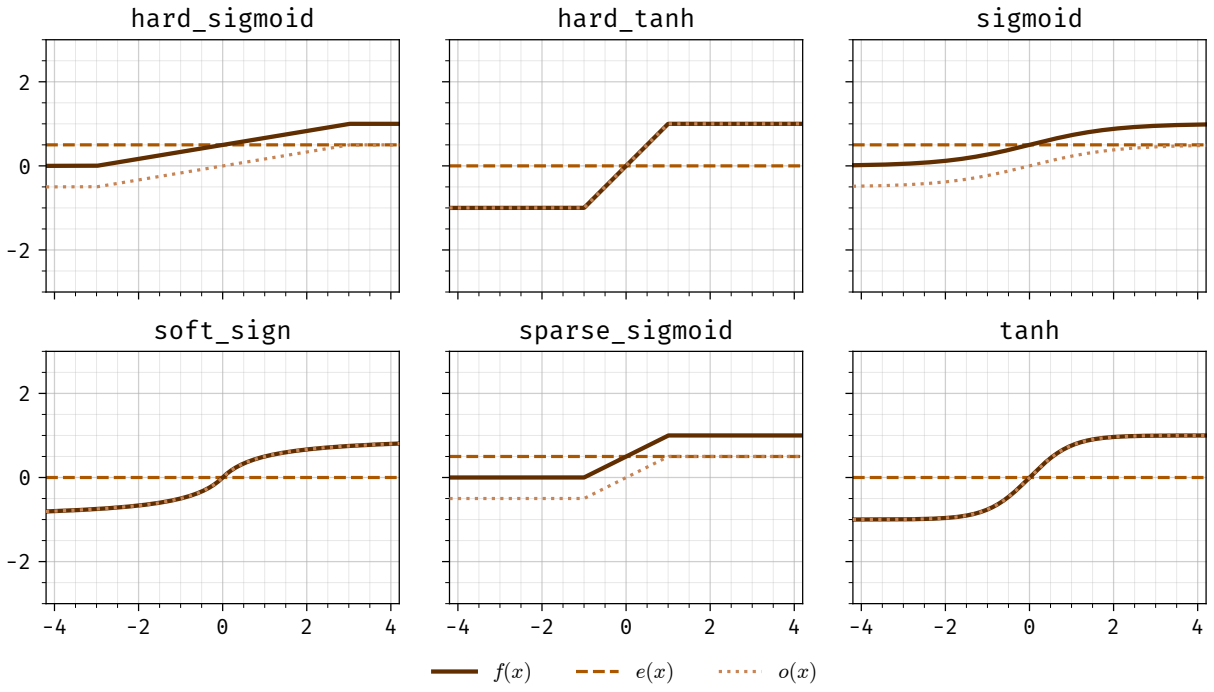

Figure F.2. Activations from `jax.nn` classified as constant-odd, shown with their decomposition into even and odd components. Solid dark lines show the activations, dashed lines their even components, and dotted lines their odd components. In each case, the even component is a horizontal line—at $y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

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

$$\sigma(-x) = \frac{1}{1+\mathrm{e}^{x}} = \frac{\mathrm{e}^{-x}}{1+\mathrm{e}^{-x}} = 1 - \frac{1}{1+\mathrm{e}^{-x}} = 1 - \sigma(x). \tag{69}$$

Hence, the even component is constant,

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{\sigma(x) + (1 - \sigma(x))}{2} = \frac{1}{2}, \tag{70}$$

so the sigmoid activation function is constant-odd with $c = \frac{1}{2}$. Applying Equation (69) once more shows that the odd component is not linear,

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = \frac{\sigma(x) - (1 - \sigma(x))}{2} = \sigma(x) - \frac{1}{2}, \tag{71}$$

so the sigmoid is not even-linear.

Softsign. The softsign activation function satisfies

$$\sigma(-x) = \frac{-x}{|-x|+1} = -\frac{x}{|x|+1} = -\sigma(x), \tag{72}$$

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

$$\sigma(x) = \begin{cases} 0, & x \le -1 \\ \frac{1}{2}(x+1), & -1 < x < 1 \\ 1, & x \ge 1 \end{cases} \quad \text{and} \quad \sigma(-x) = \begin{cases} 1, & x \le -1 \\ \frac{1}{2}(-x+1), & -1 < x < 1 \\ 0, & x \ge 1 \end{cases} \tag{73}$$

For $|x| < 1$, this gives

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{(x+1) + (-x+1)}{4} = \frac{1}{2}. \tag{74}$$

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

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = \frac{1}{2}, \qquad x \geq 1, \tag{75}$$

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

$$\tanh(x) = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}} \tag{76}$$

vanishes:

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{(\mathrm{e}^x - \mathrm{e}^{-x}) + (\mathrm{e}^{-x} - \mathrm{e}^x)}{2(\mathrm{e}^x + \mathrm{e}^{-x})} = 0. \tag{77}$$

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

F.3. Positively 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 $\sigma(\alpha x) = \alpha\sigma(x)$ for all $\alpha > 0$ and all $x \in \mathbb{R}$. The following characterization is standard:

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

$$\sigma(x) = \begin{cases} \lambda x, & x < 0 \\ \mu x, & x \geq 0 \end{cases} \tag{78}$$

Proof. Any function of the form presented in Equation (78) is readily verified to be positively homogeneous of degree 1. Conversely, suppose σ is positively homogeneous of degree 1. For $x < 0$, we have $-x > 0$ and hence

$$\sigma(x) = -x\,\sigma(-1) = \lambda x, \qquad x < 0, \tag{79}$$

with $\lambda \coloneqq -\sigma(-1)$. Similarly, for $x > 0$:

$$\sigma(x) = x\,\sigma(1) = \mu x, \qquad x > 0, \tag{80}$$

with $\mu \coloneqq \sigma(1)$. Finally, positive homogeneity implies $\sigma(0) = \sigma(\alpha \cdot 0) = \alpha\,\sigma(0)$ for every $\alpha > 0$, which requires $\sigma(0) = 0 = \mu \cdot 0$. □

The next result follows immediately:

Corollary F.2 (Unboundedness of positively homogeneous functions)**.** *On each of the half-lines $(-\infty, 0]$ and $[0, \infty)$, considered separately, a positively homogeneous function of degree 1 is either identically zero or unbounded on that half-line.*

Corollary F.2 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 F.1, and hence are positively homogeneous of degree 1.

F.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

$$\sigma(x) = \begin{cases} \alpha(\mathrm{e}^{x/\alpha} - 1), & x < 0 \\ x, & x \geq 0 \end{cases} \quad \text{and} \quad \sigma(-x) = \begin{cases} -x, & x < 0 \\ \alpha(\mathrm{e}^{-x/\alpha} - 1), & x \geq 0 \end{cases} \tag{81}$$

For $x \geq 0$, this yields

$$e(x) = \frac{x + \alpha(\mathrm{e}^{-x/\alpha} - 1)}{2} \quad \text{and} \quad o(x) = \frac{x - \alpha(\mathrm{e}^{-x/\alpha} - 1)}{2}. \tag{82}$$

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

$$\sigma(x) = \begin{cases} \alpha(\mathrm{e}^{x} - 1), & x < 0 \\ x, & x \geq 0 \end{cases} \quad \text{and} \quad \sigma(-x) = \begin{cases} -x, & x < 0 \\ \alpha(\mathrm{e}^{-x} - 1), & x \geq 0 \end{cases} \tag{83}$$

For $x \geq 0$, this yields

$$e(x) = \frac{x + \alpha(\mathrm{e}^{-x} - 1)}{2} \quad \text{and} \quad o(x) = \frac{x - \alpha(\mathrm{e}^{-x} - 1)}{2}. \tag{84}$$

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. From Equation (58), it follows

$$e(x) = \frac{\sigma(x) + \sigma(-x)}{2} = \frac{x}{2}\big(\tanh(\mathrm{softplus}(x)) - \tanh(\mathrm{softplus}(x) - x)\big) \tag{85}$$

and

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = \frac{x}{2}\big(\tanh(\mathrm{softplus}(x)) + \tanh(\mathrm{softplus}(x) - x)\big). \tag{86}$$

The even component $e(x)$ is not constant, so Mish is not constant-odd. The odd component $o(x)$ is not linear due to the presence of the $\tanh$ terms, so Mish is not even-linear.

ReLU6. From the definition of ReLU6, we have

$$\sigma(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x < 6 \\ 6, & x \geq 6 \end{cases} \quad \text{and} \quad \sigma(-x) = \begin{cases} 6, & x \leq -6 \\ -x, & -6 < x < 0 \\ 0, & x \geq 0 \end{cases} \tag{87}$$

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

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

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

$$o(x) = \frac{\sigma(x) - \sigma(-x)}{2} = 3, \qquad x \geq 6. \tag{89}$$

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

SELU. From the definition of SELU, we have

$$\sigma(x) = \lambda \begin{cases} \alpha(\mathrm{e}^{x} - 1), & x < 0 \\ x, & x \geq 0 \end{cases} \quad \text{and} \quad \sigma(-x) = \lambda \begin{cases} -x, & x < 0 \\ \alpha(\mathrm{e}^{-x} - 1), & x \geq 0 \end{cases} \tag{90}$$

For $x \geq 0$, this yields

$$e(x) = \lambda\frac{x + \alpha(\mathrm{e}^{-x} - 1)}{2} \quad \text{and} \quad o(x) = \lambda\frac{x - \alpha(\mathrm{e}^{-x} - 1)}{2}. \tag{91}$$

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.

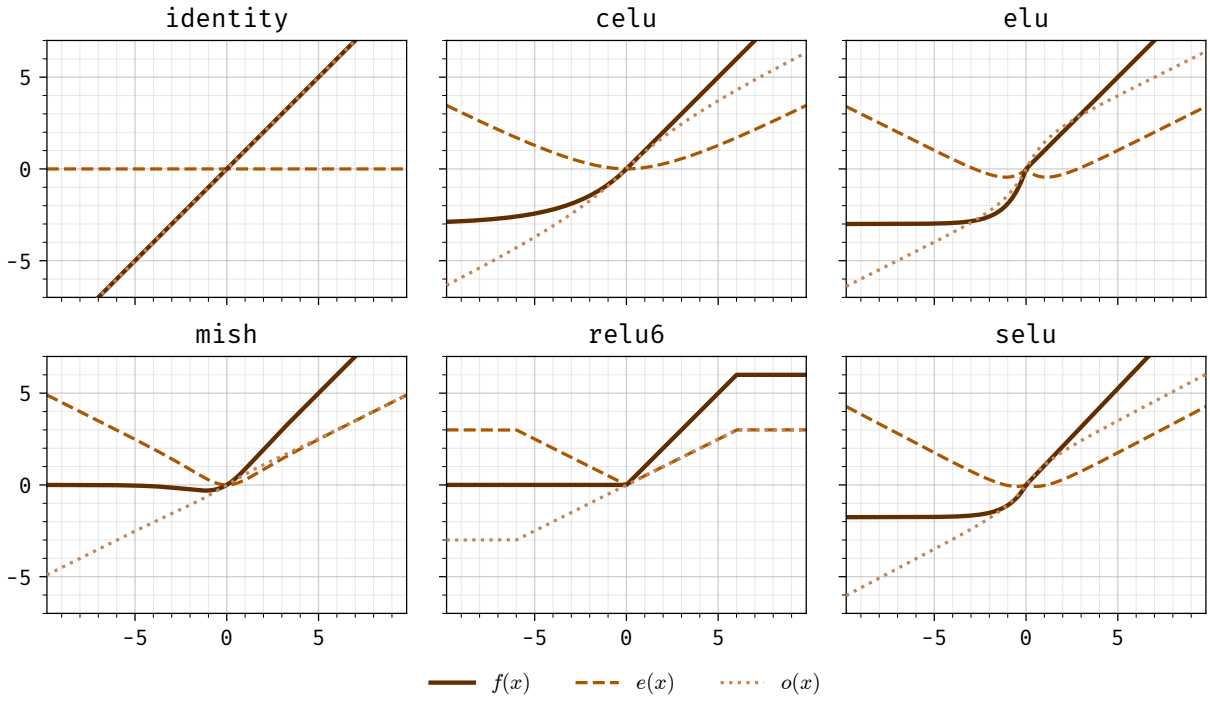

Figure F.3. Activations from `jax.nn` that are either both even-linear and constant-odd (the identity) or neither (the asymmetric activations), shown with their decomposition into even and odd components. Solid dark lines show the activations, dashed lines their even components, and dotted lines their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

G. Analytical XOR ReLU solutions

This appendix provides mathematical details supporting the ReLU networks illustrated in Figure 1. We first specify the XOR dataset, the two-hidden-unit ReLU architecture, and the binary cross-entropy 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 parametrization (Appendix G.3), and prove that each can approach zero binary cross-entropy (BCE) loss in the limit of parameter rescaling (Appendix G.4). All experiments were implemented in JAX (v0.10.0, Bradbury et al., 2018) using the Equinox neural-network library (v0.13.8, Kidger and Garcia, 2021).

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:

$$\mathbf{X} = \begin{bmatrix} -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}. \tag{92}$$

We consider two-hidden-unit ReLU networks with scalar output,

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \coloneqq a_1\, \sigma(\mathbf{w}_1^\top \mathbf{x} + b_1) + a_2\, \sigma(\mathbf{w}_2^\top \mathbf{x} + b_2) + b, \qquad \sigma(z) = \max(z, 0), \tag{93}$$

where $\mathbf{w}_i \in \mathbb{R}^2$, $b_i \in \mathbb{R}$ are the incoming weights and biases, $a_i \in \mathbb{R}$ are the readout weights, $b \in \mathbb{R}$ is the output bias, and $\boldsymbol{\theta} = (\mathbf{w}_1, b_1, \mathbf{w}_2, b_2, a_1, a_2, b)$ collects all network parameters. Training minimizes the mean binary cross-entropy (BCE) loss

$$\mathcal{L}(\boldsymbol{\theta}) \coloneqq -\frac{1}{4} \sum_{\mu=1}^{4} \left[y^\mu \log p^\mu + (1 - y^\mu) \log(1 - p^\mu) \right], \qquad p^\mu = \varsigma(f_{\boldsymbol{\theta}}(\mathbf{x}^\mu)), \tag{94}$$

where $\varsigma(x) \coloneqq (1 + \mathrm{e}^{-x})^{-1}$ denotes the logistic sigmoid.

Geometric parameterization. Each hidden neuron admits a geometric parametrization in terms of the angle ϕ_i and signed distance d_i of its activation boundary, together with a gain factor $g_i > 0$. Concretely, the pre-activation of the ith neuron is

$$z_i(\mathbf{x}) \coloneqq g_i\big(\hat{\mathbf{n}}(\phi_i)^\top \mathbf{x} - d_i\big), \qquad \hat{\mathbf{n}}(\phi) \coloneqq (\cos\phi, \sin\phi)^\top, \tag{95}$$

so that the zero level set $z_i(\mathbf{x}) = 0$ is the line with unit normal $\hat{\mathbf{n}}(\phi_i)$ at signed distance d_i from the origin. This corresponds to the usual affine parametrization via $\mathbf{w}_i = g_i\hat{\mathbf{n}}(\phi_i)$ and $b_i = -g_i d_i$.

G.2. Gradient-descent sweep

We trained `N_SEEDS` $= 1000$ two-hidden-unit 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 Equation (94). 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 Pro with an Apple M3 Max chip, 16 CPU cores, and 64 GB unified memory; it required approximately 14 minutes of wall-clock time and used at most 38 GB of peak process memory.

For each converged network, we mapped the affine hidden parameters to the geometric parametrization in Equation (95) by inverting $\mathbf{w}_i = g_i\hat{\mathbf{n}}(\phi_i)$, $b_i = -g_i d_i$, namely $g_i = \|\mathbf{w}_i\|$, $\phi_i = \operatorname{atan2}(w_{i,2}, w_{i,1})$, $d_i = -b_i/g_i$. Two structural symmetries were quotiented before clustering. The positive-scaling symmetry inherent to ReLU was removed by absorbing the gain into the readout to form effective readout weights $\beta_i = a_i g_i$. The hidden-unit permutation symmetry was then removed by sorting the two neurons lexicographically on the canonical per-neuron tuple $(\cos\phi_i, \sin\phi_i, d_i, \beta_i)$. Throughout, angles were embedded as unit-circle coordinates to avoid branch-cut artifacts at $\phi = \pm\pi$. The resulting per-network feature vector

$$(\cos\phi_1, \sin\phi_1, d_1, \beta_1, \cos\phi_2, \sin\phi_2, d_2, \beta_2, b) \in \mathbb{R}^9 \tag{96}$$

is invariant under both symmetries.

Each of the nine feature dimensions was then standardized to zero mean and unit variance across the converged set, necessary because the raw features span different scales, with the unit-circle angle coordinates bounded in $[-1, 1]$ while distances and effective readout weights take wider values, and Ward's hierarchical agglomerative clustering was applied with Euclidean distance on the standardized features. The cut threshold was placed at the largest gap in the dendrogram's merge-distance sequence. Within each resulting cluster, we selected a small set of diverse representatives by centroid-seeded farthest-first (maxmin) sampling for visualization.

Cutting the dendrogram at the largest merge-distance gap yielded six dominant clusters of sizes $88, 83, 32, 31, 31, 24$, plus a small residual cluster of 7 networks. Visual inspection of the residual cluster's members shows that each network's geometry matches that of one of the six dominant clusters. These six clusters, illustrated by representatives in Figure 1 B, fall into two families of three solutions each, which we refer to as *diagonal* and *anti-diagonal* based on the orientation of their activation boundaries. These six solution types serve as concrete examples of solutions surfaced by the sweep; we make no claim that they exhaustively classify the set of two-ReLU XOR solutions reachable by gradient descent under different optimizers, learning rates, or initializations.

G.3. Analytical form of the six solutions

Each of the six solutions found by gradient descent can be written in closed form using the geometric parametrization introduced in Equation (95). All six use unit gain ($g_1 = g_2 = 1$), so that $\mathbf{w}_i = \hat{\mathbf{n}}(\phi_i)$ and $b_i = -d_i$. Table G.1 lists the complete specifications.

Reduction to one dimension. A key structural property shared by all six solutions is that both weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are collinear. Specifically:

- Solutions 1–3 have angles $\phi_i \in \{3\pi/4, 7\pi/4\}$. Since $\hat{\mathbf{n}}(7\pi/4) = -\hat{\mathbf{n}}(3\pi/4)$, both weight vectors lie on the line spanned by $\hat{\mathbf{n}}(3\pi/4)$. The network output depends on $\mathbf{x}$ only through the *diagonal* projection $u_d = \hat{\mathbf{n}}(3\pi/4)^\top \mathbf{x} = (x_2 - x_1)/\sqrt{2}$.
- Solutions 4–6 use angles $\phi_i \in \{\pi/4, 5\pi/4\}$, and the output depends only on the *anti-diagonal* projection $u_a = \hat{\mathbf{n}}(\pi/4)^\top \mathbf{x} = (x_1 + x_2)/\sqrt{2}$.

Table G.1. The six representative ReLU solutions for XOR identified through a 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	$(\frac{5\pi}{4}, \frac{5\pi}{4})$	$(0, -\sqrt{2})$	$(-2, 1)$	$-\frac{1}{\sqrt{2}}$

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 label-0 at $u_a = \pm\sqrt{2}$ and label-1 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:

$$f(\mathbf{x}^\mu) = \begin{cases} +\frac{1}{\sqrt{2}}, & y^\mu = 1 \\ -\frac{1}{\sqrt{2}}, & y^\mu = 0 \end{cases} \qquad \mu = 1, \dots, 4. \tag{97}$$

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

$$f(\mathbf{x}) = \sigma(u_d) + \sigma(-u_d) - \frac{1}{\sqrt{2}} = |u_d| - \frac{1}{\sqrt{2}}, \tag{98}$$

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

$$f(\mathbf{x}) = -\sigma(u_a) - \sigma(-u_a) + \frac{1}{\sqrt{2}} = -|u_a| + \frac{1}{\sqrt{2}}, \tag{99}$$

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 via 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 $\boldsymbol{\theta}^\star$ be any of the six solutions from Table G.1, and define the scaled parameter vector*

$$\boldsymbol{\theta}^\star_\alpha := (\alpha \mathbf{w}_1, \alpha b_1, \alpha \mathbf{w}_2, \alpha b_2, a_1, a_2, \alpha b), \qquad \alpha > 0. \tag{100}$$

Then $f_{\boldsymbol{\theta}^\star_\alpha}(\mathbf{x}^\mu) = \alpha\, f_{\boldsymbol{\theta}^\star}(\mathbf{x}^\mu)$, for $\mu = 1, \dots, 4$, and the BCE loss along the ray generated by α satisfies

$$\mathcal{L}(\alpha) := \mathcal{L}(\boldsymbol{\theta}^\star_\alpha) = \log(1 + \exp(-\alpha/\sqrt{2})), \tag{101}$$

which is strictly decreasing in α, with $\lim_{\alpha\to\infty} \mathcal{L}(\alpha) = 0$.

Proof. Under the scaling in Equation (100), the network output satisfies $f_{\boldsymbol{\theta}^\star_\alpha}(\mathbf{x}) = \alpha\, f_{\boldsymbol{\theta}^\star}(\mathbf{x})$, since ReLU is positively homogeneous of degree 1. By Equation (97), the logit at each data point thus has magnitude $\alpha/\sqrt{2}$ with the appropriate

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

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

sign. For a label-1 point, the BCE contribution is $-\log \varsigma(\alpha/\sqrt{2})$. For a label-0 point, it is $-\log(1-\varsigma(-\alpha/\sqrt{2})) = -\log \varsigma(\alpha/\sqrt{2})$, 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 $\mathcal{L}(\alpha) = \log(1+\exp(-\alpha/\sqrt{2}))$. The map $\alpha \mapsto -\alpha/\sqrt{2}$ is strictly decreasing, while exp and log are strictly increasing; hence $\mathcal{L}(\alpha)$ is strictly decreasing. The fact that $\lim_{\alpha\to\infty} \mathcal{L}(\alpha) = 0$ is evident. □

H. Proofs

This appendix contains the proofs and technical details deferred from the main text. Its organization mirrors that of the paper: each appendix section corresponds to a section of the main text and collects the associated proofs, auxiliary arguments, and additional details.

H.1. Proofs for Section 4

This subsection contains the proofs for the results in Section 4. We begin by introducing *duplication patterns* and the *duplication matrices* through which they act, along with the elementary identities governing their multiplication (Appendix H.1.1). We then prove that each parameter symmetry cataloged in Appendix E induces, at the level of hidden activations, a finite composition of feature additions, duplications, and scalings (Appendix H.1.2). Next, we establish the elementary identities describing how these primitive feature transforms commute past one another and collapse pairwise into single primitives of the same type (Appendix H.1.3). Finally, we use these identities to reduce any finite composition of primitive feature transforms to the canonical add–duplicate–scale form stated in Proposition 4.2 (Appendix H.1.4).

H.1.1. Duplication matrices

Definition H.1 (Duplication pattern)**.** Let $N_h^\star \in \mathbb{N}$ denote the width of a hidden layer. A *duplication pattern* is a vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{N_h^\star})^\top$, where $\nu_i \in \mathbb{N}_{>0}$ denotes the number of copies of the ith neuron after duplication. We refer to the resulting width of the transformed layer, given by

$$N_h \coloneqq \sum_{i=1}^{N_h^\star} \nu_i, \tag{102}$$

as the duplication pattern's *induced width*.

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

$$\mathbf{D}_{\boldsymbol{\nu}} \coloneqq \begin{bmatrix} \mathbf{1}_{\nu_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{\nu_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{\nu_{N_h^\star}} \end{bmatrix} \tag{103}$$

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

$$(\mathbf{D}_{\boldsymbol{\nu}})_{rj} \coloneqq \begin{cases} 1, & \text{if } \sum_{i=1}^{j-1} \nu_i < r \le \sum_{i=1}^{j} \nu_i \\ 0, & \text{otherwise} \end{cases} \tag{104}$$

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

$$\nu_{<j} \coloneqq \sum_{i=1}^{j-1} \nu_i, \qquad j = 1, \ldots, N_h^\star, \tag{105}$$

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 (104) as

$$(\mathbf{D}_{\boldsymbol{\nu}})_{rj} = \begin{cases} 1, & \text{if } \nu_{<j} < r \leq \nu_{<(j+1)} \\ 0, & \text{otherwise} \end{cases} \tag{106}$$

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 H.1 and made precise by the following result.

Lemma H.4 (Multiplication by duplication matrices)**.** *Let $\boldsymbol{\nu} \in \mathbb{N}_{>0}^{N_h^\star}$ be a duplication pattern with induced width N_h. Further, let $\mathbf{q} \in \mathbb{R}^{N_h^\star}$, let $\mathbf{M} \in \mathbb{R}^{N_h^\star \times Q}$, and let $\mathbf{N} \in \mathbb{R}^{Q \times N_h^\star}$. Then*

1. $(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{q})_r = q_j$
2. $(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{M})_{r,:} = \mathbf{M}_{j,:}$
3. $(\mathbf{N}\mathbf{D}_{\boldsymbol{\nu}}^\top)_{:,r} = \mathbf{N}_{:,j}$

if and only if $\nu_{<j} < r \leq \nu_{<(j+1)}$.

Proof. For part (i), we have

$$(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{q})_r = \sum_{i=1}^{N_h^\star} (\mathbf{D}_{\boldsymbol{\nu}})_{ri} q_i. \tag{107}$$

Since the intervals $(\nu_{<i}, \nu_{<(i+1)}]$ partition $\{1, \ldots, N_h\}$, each r 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 $(\mathbf{D}_{\boldsymbol{\nu}}\mathbf{q})_r = q_j$. Part (ii) follows by applying (i) to each column of $\mathbf{M}$, and part (iii) follows by transposing (ii). □

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

$$\mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\gamma}) \mathbf{D}_{\boldsymbol{\nu}} = \operatorname{diag}(\mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\gamma}). \tag{108}$$

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 $\mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\gamma}) \mathbf{D}_{\boldsymbol{\nu}}$ equals

$$(\mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\gamma}) \mathbf{D}_{\boldsymbol{\nu}})_{rj} = \mathbf{d}_r^\top \operatorname{diag}(\boldsymbol{\gamma}) \mathbf{d}_j = \sum_{i=1}^{N_h} \gamma_i (\mathbf{d}_r)_i (\mathbf{d}_j)_i. \tag{109}$$

This sum accumulates the weights γ_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

$$(\mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\gamma}) \mathbf{D}_{\boldsymbol{\nu}})_{jj} = \sum_{i=1}^{N_h} \gamma_i (\mathbf{d}_j)_i^2 = \sum_{i=1}^{N_h} \gamma_i (\mathbf{d}_j)_i = \mathbf{d}_j^\top \boldsymbol{\gamma} = (\mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\gamma})_j, \tag{110}$$

completing the proof. □

Table H.2. Decomposition of parameter symmetries into the three primitive feature transforms. Checkmarks indicate which of the feature transforms—addition, duplication, and scaling—appear in each symmetry's decomposition; the absence of all three (top row) corresponds to the trivial action of permuting the rows of the hidden activation matrix $\mathbf{H}$, which leaves the Gram matrix $\mathbf{H}^\top\mathbf{H}$ invariant.

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	✓	✓	✗

H.1.2. Primitive feature transforms generate hidden representations within orbits

Proof of Proposition 4.1. By definition, the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ consists of all hidden-layer parameterizations $\boldsymbol{\theta}$ of width $N_h \geq N_h^\star$ obtained from the irreducible parameterization $\boldsymbol{\theta}^\star$ by a finite composition of the parameter symmetries cataloged in Appendix E. It therefore suffices to show that each cataloged parameter symmetry induces, at the level of hidden activations, a finite composition of feature additions, duplications, and scalings. Indeed, any finite composition of parameter symmetries then induces a finite composition of the corresponding primitive feature transformations. We now verify this claim for the symmetries in Appendix E, following the order in which they are introduced there.

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

$$\mathbf{H} = \mathrm{diag}(\boldsymbol{\alpha})\mathbf{H}^\star, \qquad \alpha_i = \begin{cases} 1, & i \neq j \\ \alpha_j, & i = j \end{cases} \tag{111}$$

Sign-flip symmetry. For odd activations, flipping the signs of the jth neuron's incoming weights $\mathbf{w}_j^\star$, bias $b_j^\star$, and readout weights $\mathbf{a}_j^\star$ leaves the represented function unchanged. Since $\sigma(-x) = -\sigma(x)$, the hidden feature computed by the jth neuron changes sign, so the jth row of $\mathbf{H}^\star$ is multiplied by -1. This is an instance of feature scaling acting on the jth row:

$$\mathbf{H} = \mathrm{diag}(\boldsymbol{\alpha})\mathbf{H}^\star, \qquad \alpha_i = \begin{cases} 1, & i \neq j \\ -1, & i = j \end{cases} \tag{112}$$

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

$$\mathbf{H} = \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{u} \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} \mathbf{1} \\ K \end{bmatrix}. \tag{113}$$

Duplicate-neuron groups. Replacing the jth neuron with a duplicate-neuron group of size K replicates the jth row of $\mathbf{H}^\star$ into K copies. This is an instance of feature duplication acting on the jth row:

$$\mathbf{H} = \mathbf{D}_{\boldsymbol{\nu}}\mathbf{H}^\star, \qquad \nu_i = \begin{cases} 1, & i \neq j \\ K, & i = j \end{cases} \tag{114}$$

Constant neurons. Since a constant neuron has zero incoming weights, its activation $\sigma(b)$ 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

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{u} \end{bmatrix}. \tag{115}$$

Linear-neuron groups. Since aligned and opposite subgroups of a linear-neuron group share incoming weights up to a sign flip, an even-linear activation $\sigma(x) = e(x) + mx$ yields exactly two distinct activation patterns. Writing

$$\mathbf{v} \coloneqq \mathbf{w}^\top \mathbf{X} + b\mathbf{1}^\top \in \mathbb{R}^{1\times P} \tag{116}$$

for the pre-activation feature vector, aligned neurons produce $\mathbf{u}^+ \coloneqq e(\mathbf{v}) + m\mathbf{v}$, whereas opposite neurons produce $\mathbf{u}^- \coloneqq e(\mathbf{v}) - m\mathbf{v}$, with e applied elementwise. 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:

$$\mathbf{H} = \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{u}^+ \\ \mathbf{u}^- \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} \mathbf{1} \\ |\mathcal{N}^+| \\ |\mathcal{N}^-| \end{bmatrix}. \tag{117}$$

Linear-duplicate-neuron groups. Since aligned neurons share the incoming weights and bias of the jth neuron, they replicate the jth row of $\mathbf{H}^\star$. Opposite neurons, having sign-flipped incoming weights and bias, produce a single distinct activation pattern. Writing

$$\mathbf{v}_j \coloneqq \mathbf{w}_j^{\star\top} \mathbf{X} + b_j^\star \mathbf{1}^\top \in \mathbb{R}^{1\times P} \tag{118}$$

for the pre-activation feature vector of the jth neuron, opposite neurons produce $\mathbf{u}^- \coloneqq e(\mathbf{v}_j) - m\mathbf{v}_j$. 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 $\mathbf{H}^\star$ for each aligned neuron and $\mathbf{u}^-$ for each opposite neuron:

$$\mathbf{H} = \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{u}^- \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} \boldsymbol{\nu}^+ \\ |\mathcal{N}^-| \end{bmatrix}, \qquad \nu_i^+ = \begin{cases} 1, & i \neq j \\ |\mathcal{N}^+|, & i = j \end{cases} \tag{119}$$

Constant-neuron groups. For a constant-odd activation $\sigma(x) = c + o(x)$, aligned and opposite subgroups produce $\mathbf{u}^+ \coloneqq c\mathbf{1}^\top + o(\mathbf{v})$ and $\mathbf{u}^- \coloneqq c\mathbf{1}^\top - o(\mathbf{v})$, respectively, where $\mathbf{v}$ is defined as in Equation (116). The resulting transformation is otherwise identical to Equation (117).

Constant-duplicate-neuron groups. Opposite neurons produce $\mathbf{u}^- \coloneqq c\mathbf{1}^\top - o(\mathbf{v}_j)$, where $\mathbf{v}_j$ is the pre-activation feature vector of the jth neuron defined in Equation (118). The resulting transformation is otherwise identical to Equation (119). □

H.1.3. Primitive feature transforms commute and collapse

Lemma H.6 (Primitive feature transforms commute)**.** *Let $\mathbf{H} \in \mathbb{R}^{N_h^\star \times P}$ and $\mathbf{U} \in \mathbb{R}^{K\times P}$. Further, let $\boldsymbol{\nu} \in \mathbb{N}_{>0}^{N_h^\star}$ be a duplication pattern with induced width N_h and associated duplication matrix $\mathbf{D}_{\boldsymbol{\nu}} \in \{0,1\}^{N_h \times N_h^\star}$, and let $\boldsymbol{\alpha} \in \mathbb{R}_{\neq 0}^{N_h^\star}$. Define*

$$\widetilde{\boldsymbol{\nu}} = \begin{bmatrix} \boldsymbol{\nu} \\ \mathbf{1} \end{bmatrix} \in \mathbb{N}_{>0}^{N_h^\star + K} \quad \text{and} \quad \widetilde{\boldsymbol{\alpha}} = \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{1} \end{bmatrix} \in \mathbb{R}_{\neq 0}^{N_h^\star + K}. \tag{120}$$

Then the following identities hold:

1. $\begin{bmatrix} \mathbf{D}_{\boldsymbol{\nu}}\mathbf{H} \\ \mathbf{U} \end{bmatrix} = \mathbf{D}_{\widetilde{\boldsymbol{\nu}}} \begin{bmatrix} \mathbf{H} \\ \mathbf{U} \end{bmatrix}$
2. $\begin{bmatrix} \operatorname{diag}(\boldsymbol{\alpha})\mathbf{H} \\ \mathbf{U} \end{bmatrix} = \operatorname{diag}(\widetilde{\boldsymbol{\alpha}}) \begin{bmatrix} \mathbf{H} \\ \mathbf{U} \end{bmatrix}$
3. $\mathbf{D}_{\boldsymbol{\nu}} \operatorname{diag}(\boldsymbol{\alpha}) = \operatorname{diag}(\mathbf{D}_{\boldsymbol{\nu}}\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}}$

Proof. For part (i), note that the duplication matrix $\mathbf{D}_{\widetilde{\boldsymbol{\nu}}}$ can be written as the block matrix

$$\mathbf{D}_{\widetilde{\boldsymbol{\nu}}} = \begin{bmatrix} \mathbf{D}_{\boldsymbol{\nu}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{121}$$

where $\mathbf{I}$ denotes the identity matrix of size $K \times K$. The claimed identity then follows directly from block matrix multiplication:

$$\begin{bmatrix} \mathbf{D}_{\boldsymbol{\nu}}\mathbf{H} \\ \mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{\boldsymbol{\nu}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{H} \\ \mathbf{U} \end{bmatrix} = \mathbf{D}_{\widetilde{\boldsymbol{\nu}}} \begin{bmatrix} \mathbf{H} \\ \mathbf{U} \end{bmatrix}. \tag{122}$$

Part (ii) follows analogously, as the diagonal matrix $\mathrm{diag}(\widetilde{\boldsymbol{\alpha}})$ also has a simple block structure:

$$\mathrm{diag}(\widetilde{\boldsymbol{\alpha}}) = \begin{bmatrix} \mathrm{diag}(\boldsymbol{\alpha}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{123}$$

For part (iii), since $\mathrm{diag}(\boldsymbol{\alpha})$ has nonzero entries only on its diagonal, we have

$$(\mathbf{D}_{\boldsymbol{\nu}}\,\mathrm{diag}(\boldsymbol{\alpha}))_{jr} = \sum_{i=1}^{N_h^\star} (\mathbf{D}_{\boldsymbol{\nu}})_{ji}(\mathrm{diag}(\boldsymbol{\alpha}))_{ir} = (\mathbf{D}_{\boldsymbol{\nu}})_{jr}\alpha_r = \begin{cases} \alpha_r, & \text{if } \nu_{<r} < j \le \nu_{<(r+1)} \\ 0, & \text{otherwise} \end{cases} \tag{124}$$

On the other hand,

$$(\mathrm{diag}(\mathbf{D}_{\boldsymbol{\nu}}\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}})_{jr} = \sum_{i=1}^{N_h} (\mathrm{diag}(\mathbf{D}_{\boldsymbol{\nu}}\boldsymbol{\alpha}))_{ji}(\mathbf{D}_{\boldsymbol{\nu}})_{ir} = (\mathbf{D}_{\boldsymbol{\nu}}\boldsymbol{\alpha})_j(\mathbf{D}_{\boldsymbol{\nu}})_{jr}. \tag{125}$$

By definition, $(\mathbf{D}_{\boldsymbol{\nu}})_{jr} = 1$ if and only if $\nu_{<r} < j \le \nu_{<(r+1)}$, which is equivalent to $(\mathbf{D}_{\boldsymbol{\nu}}\boldsymbol{\alpha})_j = \alpha_r$ by Lemma H.4. Therefore,

$$(\mathrm{diag}(\mathbf{D}_{\boldsymbol{\nu}}\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}})_{jr} = \begin{cases} \alpha_r, & \text{if } \nu_{<r} < j \le \nu_{<(r+1)} \\ 0, & \text{otherwise} \end{cases} \tag{126}$$

and the identity follows. □

Lemma H.7 (Primitive feature transforms collapse)**.** *The following identities hold:*

1. *Let $\mathbf{H} \in \mathbb{R}^{N_h^\star \times P}$, $\mathbf{U}^{(1)} \in \mathbb{R}^{K_1 \times P}$, and $\mathbf{U}^{(2)} \in \mathbb{R}^{K_2 \times P}$. Then*

$$\begin{bmatrix} \begin{bmatrix} \mathbf{H} \\ \mathbf{U}^{(1)} \end{bmatrix} \\ \mathbf{U}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{H} \\ \begin{bmatrix} \mathbf{U}^{(1)} \\ \mathbf{U}^{(2)} \end{bmatrix} \end{bmatrix}. \tag{127}$$

2. *Let $\boldsymbol{\nu}^{(1)} \in \mathbb{N}_{>0}^{N_h^\star}$ and $\boldsymbol{\nu}^{(2)} \in \mathbb{N}_{>0}^{N_h}$ be duplication patterns with induced widths N_h and N_h', respectively. Then*

$$\mathbf{D}_{\boldsymbol{\nu}^{(2)}}\mathbf{D}_{\boldsymbol{\nu}^{(1)}} = \mathbf{D}_{\boldsymbol{\nu}^{(2)} \circ \boldsymbol{\nu}^{(1)}}, \tag{128}$$

   where $\boldsymbol{\nu}^{(2)} \circ \boldsymbol{\nu}^{(1)} \in \mathbb{N}_{>0}^{N_h^\star}$ is the composite duplication pattern defined by

$$(\boldsymbol{\nu}^{(2)} \circ \boldsymbol{\nu}^{(1)})_j = \sum_{i=1}^{\nu_j^{(1)}} \nu^{(2)}_{\nu^{(1)}_{<j}+i}, \qquad j = 1, \ldots, N_h^\star. \tag{129}$$

3. *Let $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)} \in \mathbb{R}_{\neq 0}^{N_h^\star}$. Then*

$$\mathrm{diag}(\boldsymbol{\alpha}^{(1)})\,\mathrm{diag}(\boldsymbol{\alpha}^{(2)}) = \mathrm{diag}(\boldsymbol{\alpha}^{(1)} \odot \boldsymbol{\alpha}^{(2)}), \tag{130}$$

   where $\odot$ denotes the Hadamard product.

Proof. Part (i) is immediate from the associativity of vertical concatenation of matrices.

For part (ii), write $\boldsymbol{\nu} = \boldsymbol{\nu}^{(2)} \circ \boldsymbol{\nu}^{(1)}$ for the composite pattern defined in Equation (129). Left multiplication of any matrix $\mathbf{M} \in \mathbb{R}^{N_h^\star \times Q}$ by $\mathbf{D}_{\boldsymbol{\nu}^{(1)}}$ creates $\nu_j^{(1)}$ copies of the jth row of $\mathbf{M}$ at positions

$$\nu_{<j}^{(1)} + 1, \dots, \nu_{<j}^{(1)} + \nu_j^{(1)}, \qquad j = 1, \dots, N_h^\star. \tag{131}$$

Multiplying the resulting matrix from the left by $\mathbf{D}_{\boldsymbol{\nu}^{(2)}}$ then creates $\nu_i^{(2)}$ copies of row i at positions

$$\nu_{<i}^{(2)} + 1, \dots, \nu_{<i}^{(2)} + \nu_i^{(2)}, \qquad i = \nu_{<j}^{(1)} + 1, \dots, \nu_{<j}^{(1)} + \nu_j^{(1)}. \tag{132}$$

Thus, the product $\mathbf{D}_{\boldsymbol{\nu}^{(2)}} \mathbf{D}_{\boldsymbol{\nu}^{(1)}}$ produces a total of

$$\sum_{i=\nu_{<j}^{(1)}+1}^{\nu_{<j}^{(1)}+\nu_j^{(1)}} \nu_i^{(2)} = \sum_{i=1}^{\nu_j^{(1)}} \nu_{\nu_{<j}^{(1)}+i}^{(2)} = \nu_j \tag{133}$$

copies of the jth row of $\mathbf{M}$, placed at positions

$$\nu_{<(\nu_{<j}^{(1)}+1)}^{(2)} + 1, \dots, \nu_{<(\nu_{<j}^{(1)}+1)}^{(2)} + \nu_j, \qquad j = 1, \dots, N_h^\star. \tag{134}$$

It remains to verify that these positions match those produced by $\mathbf{D}_{\boldsymbol{\nu}}$. To this end, we rewrite the offset

$$\nu_{<(\nu_{<j}^{(1)}+1)}^{(2)} = \sum_{i=1}^{\nu_{<j}^{(1)}} \nu_i^{(2)} \tag{135}$$

determining the positions in Equation (134) of the copies corresponding to the jth row. The index set $\{1, \dots, \nu_{<j}^{(1)}\}$ can be partitioned into disjoint consecutive blocks

$$\{\nu_{<p}^{(1)} + 1, \dots, \nu_{<p}^{(1)} + \nu_p^{(1)}\}, \qquad p = 1, \dots, j-1. \tag{136}$$

Using this partition, we obtain

$$\sum_{i=1}^{\nu_{<j}^{(1)}} \nu_i^{(2)} = \sum_{p=1}^{j-1} \sum_{i=\nu_{<p}^{(1)}+1}^{\nu_{<p}^{(1)}+\nu_p^{(1)}} \nu_i^{(2)} = \sum_{p=1}^{j-1} \sum_{i=1}^{\nu_p^{(1)}} \nu_{\nu_{<p}^{(1)}+i}^{(2)} = \sum_{p=1}^{j-1} \nu_p = \nu_{<j}. \tag{137}$$

Consequently, the ν_j copies of the jth row produced by $\mathbf{D}_{\boldsymbol{\nu}^{(2)}} \mathbf{D}_{\boldsymbol{\nu}^{(1)}}$ occupy exactly the positions

$$\nu_{<j} + 1, \dots, \nu_{<j} + \nu_j, \qquad j = 1, \dots, N_h^\star, \tag{138}$$

which coincides with the action of left multiplication by $\mathbf{D}_{\boldsymbol{\nu}}$.

Part (iii) follows from the fact that the product of two diagonal matrices is another diagonal matrix with entries

$$(\operatorname{diag}(\boldsymbol{\alpha}^{(1)}) \operatorname{diag}(\boldsymbol{\alpha}^{(2)}))_{jj} = \alpha_j^{(1)} \alpha_j^{(2)} = (\boldsymbol{\alpha}^{(1)} \odot \boldsymbol{\alpha}^{(2)})_j, \qquad j = 1, \dots, N_h^\star, \tag{139}$$

on the diagonal. □

H.1.4. Deriving a canonical form of the hidden activation matrix

Proof of Proposition 4.2. Proposition 4.1 guarantees that the hidden activation matrix $\mathbf{H}$ induced by a layer $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ can be constructed from the hidden activation matrix $\mathbf{H}^\star$ of the irreducible representative $\boldsymbol{\theta}^\star$ by a finite composition of feature additions, duplications, and scalings. Here, we show that any such sequence of primitive transforms can be reordered and collapsed to yield the canonical form

$$\mathbf{H} = \operatorname{diag}(\boldsymbol{\alpha}) \mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{U} \end{bmatrix}, \qquad \boldsymbol{\alpha} \in \mathbb{R}_{\neq 0}^{N_h}, \quad \boldsymbol{\nu} \in \mathbb{N}_{>0}^{N_h^\star + K}, \quad \mathbf{U} \in \mathbb{R}^{K \times P}, \tag{7}$$

restated here for convenience. Let $\mathcal{A}_{\mathbf{U}}$, $\mathcal{D}_{\boldsymbol{\nu}}$, and $\mathcal{S}_{\boldsymbol{\alpha}}$ denote the operators corresponding to feature addition of $\mathbf{U}$, feature duplication according to $\mathbf{D}_{\boldsymbol{\nu}}$, and feature scaling by $\mathrm{diag}(\boldsymbol{\alpha})$, respectively. That is,

$$\mathcal{A}_{\mathbf{U}}(\mathbf{M}) \coloneqq \begin{bmatrix}\mathbf{M}\\ \mathbf{U}\end{bmatrix}, \qquad \mathcal{D}_{\boldsymbol{\nu}}(\mathbf{M}) \coloneqq \mathbf{D}_{\boldsymbol{\nu}}\mathbf{M}, \qquad \mathcal{S}_{\boldsymbol{\alpha}}(\mathbf{M}) \coloneqq \mathrm{diag}(\boldsymbol{\alpha})\mathbf{M}. \tag{140}$$

Proposition 4.1 implies that

$$\mathbf{H} = (\mathcal{T}_p \circ \cdots \circ \mathcal{T}_1)(\mathbf{H}^\star), \tag{141}$$

for a finite sequence of primitive transformations

$$\mathcal{T}_i \in \{\mathcal{A}_{\mathbf{U}^{(i)}}, \mathcal{D}_{\boldsymbol{\nu}^{(i)}}, \mathcal{S}_{\boldsymbol{\alpha}^{(i)}}\}, \qquad i = 1, \ldots, p. \tag{142}$$

Step 1: Reordering. We first show that the composition in Equation (141) can be reordered so that all feature additions $\mathcal{A}_{\mathbf{U}^{(i)}}$ are performed first, followed by duplications $\mathcal{D}_{\boldsymbol{\nu}^{(i)}}$ and scalings $\mathcal{S}_{\boldsymbol{\alpha}^{(i)}}$, in that order. Identities (i) and (ii) of Lemma H.6 translate to

$$\mathcal{A}_{\mathbf{U}} \circ \mathcal{D}_{\boldsymbol{\nu}} = \mathcal{D}_{\widetilde{\boldsymbol{\nu}}} \circ \mathcal{A}_{\mathbf{U}} \qquad \text{and} \qquad \mathcal{A}_{\mathbf{U}} \circ \mathcal{S}_{\boldsymbol{\alpha}} = \mathcal{S}_{\widetilde{\boldsymbol{\alpha}}} \circ \mathcal{A}_{\mathbf{U}}, \tag{143}$$

where $\widetilde{\boldsymbol{\nu}}$ and $\widetilde{\boldsymbol{\alpha}}$ are the extended parameters defined in Equation (120). Applying these identities repeatedly, swapping an addition past a neighboring duplication or scaling at each step, gives

$$\mathbf{H} = (\widetilde{\mathcal{T}}_p \circ \cdots \circ \widetilde{\mathcal{T}}_{a+1}) \circ (\mathcal{A}_{\mathbf{U}^{(a)}} \circ \cdots \circ \mathcal{A}_{\mathbf{U}^{(1)}})(\mathbf{H}^\star), \tag{144}$$

where each

$$\widetilde{\mathcal{T}}_i \in \{\mathcal{D}_{\widetilde{\boldsymbol{\nu}}^{(i)}}, \mathcal{S}_{\widetilde{\boldsymbol{\alpha}}^{(i)}}\}, \qquad i = a+1, \ldots, p \tag{145}$$

is either a duplication or a scaling, with parameters updated according to the commutativity rules stated in Equation (143). Next, identity (iii) of Lemma H.6 implies

$$\mathcal{D}_{\boldsymbol{\nu}} \circ \mathcal{S}_{\boldsymbol{\alpha}} = \mathcal{S}_{\mathbf{D}_{\boldsymbol{\nu}}\boldsymbol{\alpha}} \circ \mathcal{D}_{\boldsymbol{\nu}}. \tag{146}$$

Consequently, we can iteratively swap any duplication past a neighboring scaling to further transform Equation (144) into

$$\mathbf{H} = (\mathcal{S}_{\boldsymbol{\alpha}^{(s)}} \circ \cdots \circ \mathcal{S}_{\boldsymbol{\alpha}^{(1)}}) \circ (\mathcal{D}_{\boldsymbol{\nu}^{(d)}} \circ \cdots \circ \mathcal{D}_{\boldsymbol{\nu}^{(1)}}) \circ (\mathcal{A}_{\mathbf{U}^{(a)}} \circ \cdots \circ \mathcal{A}_{\mathbf{U}^{(1)}})(\mathbf{H}^\star), \tag{147}$$

where scaling parameters $\boldsymbol{\alpha}^{(i)}$ are updated according to Equation (146).

Step 2: Reduction. Finally, we combine multiple primitive transformations of the same type into a single primitive of the corresponding type. For brevity, denote the vertical concatenation of $\mathbf{U}^{(i)}, \ldots, \mathbf{U}^{(j)}$ by

$$\mathbf{U}^{(i:j)} = \begin{bmatrix}\mathbf{U}^{(i)}\\ \vdots \\ \mathbf{U}^{(j)}\end{bmatrix}. \tag{148}$$

By Lemma H.7,

$$\mathcal{A}_{\mathbf{U}^{(2)}} \circ \mathcal{A}_{\mathbf{U}^{(1)}} = \mathcal{A}_{\mathbf{U}^{(1:2)}}, \quad \mathcal{D}_{\boldsymbol{\nu}^{(2)}} \circ \mathcal{D}_{\boldsymbol{\nu}^{(1)}} = \mathcal{D}_{\boldsymbol{\nu}^{(2)} \circ \boldsymbol{\nu}^{(1)}}, \quad \mathcal{S}_{\boldsymbol{\alpha}^{(2)}} \circ \mathcal{S}_{\boldsymbol{\alpha}^{(1)}} = \mathcal{S}_{\boldsymbol{\alpha}^{(1)} \odot \boldsymbol{\alpha}^{(2)}}. \tag{149}$$

Therefore, Equation (147) simplifies to

$$\mathbf{H} = (\mathcal{S}_{\boldsymbol{\alpha}} \circ \mathcal{D}_{\boldsymbol{\nu}} \circ \mathcal{A}_{\mathbf{U}})(\mathbf{H}^\star) = \mathrm{diag}(\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix}\mathbf{H}^\star\\ \mathbf{U}\end{bmatrix} \tag{150}$$

for

$$\mathbf{U} = \mathbf{U}^{(1:a)}, \quad \boldsymbol{\nu} = \boldsymbol{\nu}^{(d)} \circ \cdots \circ \boldsymbol{\nu}^{(1)}, \quad \boldsymbol{\alpha} = \boldsymbol{\alpha}^{(1)} \odot \cdots \odot \boldsymbol{\alpha}^{(s)}, \tag{151}$$

which is exactly Equation (7). □

H.2. Proofs for Section B

This subsection contains the proofs for the results in Section B. We first decompose the RSM of any orbit element into rank-one contributions from irreducible and symmetry-induced features (Appendix H.2.1). We then use this decomposition to characterize the set of attainable representational geometries through a sequence of nested convex cones (Appendix H.2.2). Finally, we establish that the freedom captured by these cones translates into a closed interval of representational similarity scores against any fixed reference geometry, with the interval growing as overparameterization admits further symmetry-induced features (Appendix H.2.3).

H.2.1. Decomposing RSMs into task-linked and symmetry-induced components

Proof of Proposition B.1. We write

$$\mathbf{M} \coloneqq \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{U} \end{bmatrix} \in \mathbb{R}^{(N_h^\star + K) \times P}, \tag{152}$$

so that the hidden activation matrix $\mathbf{H}$ induced by $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ has canonical form

$$\mathbf{H} = \operatorname{diag}(\boldsymbol{\alpha}) \mathbf{D}_{\boldsymbol{\nu}} \mathbf{M} \tag{153}$$

by Proposition 4.2. This gives

$$\mathrm{RSM} = \mathbf{H}^\top \mathbf{H} = \mathbf{M}^\top \mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\alpha})^2 \mathbf{D}_{\boldsymbol{\nu}} \mathbf{M}. \tag{154}$$

By Lemma H.5, we have

$$\mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\alpha})^2 \mathbf{D}_{\boldsymbol{\nu}} = \mathbf{D}_{\boldsymbol{\nu}}^\top \operatorname{diag}(\boldsymbol{\alpha}^2) \mathbf{D}_{\boldsymbol{\nu}} = \operatorname{diag}(\mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2), \tag{155}$$

where $\boldsymbol{\alpha}^2$ denotes the Hadamard square of $\boldsymbol{\alpha}$. Substituting into Equation (154) yields

$$\mathrm{RSM} = \mathbf{M}^\top \operatorname{diag}(\mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2) \mathbf{M} = \sum_{i=1}^{N_h^\star + K} \gamma_i \, \mathbf{m}_i \mathbf{m}_i^\top, \qquad \boldsymbol{\gamma} = \mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2 \in \mathbb{R}_{>0}^{N_h^\star + K}, \tag{156}$$

where $\mathbf{m}_i^\top$ denotes the ith row of $\mathbf{M}$. Since

$$\mathbf{m}_i^\top = \begin{cases} \mathbf{z}_i^\top, & i \le N_h^\star \\ \mathbf{u}_{i - N_h^\star}^\top, & i > N_h^\star \end{cases} \tag{157}$$

where $\mathbf{z}_\ell^\top$ and $\mathbf{u}_\ell^\top$ denote the ℓth row of $\mathbf{H}^\star$ and $\mathbf{U}$, respectively, Equation (156) can be rewritten as

$$\mathrm{RSM} = \sum_{j=1}^{N_h^\star} \gamma_j \, \mathbf{z}_j \mathbf{z}_j^\top + \sum_{k=1}^{K} \gamma_{N_h^\star + k} \, \mathbf{u}_k \mathbf{u}_k^\top, \tag{158}$$

which is precisely Equation (8). □

H.2.2. Characterizing representational geometries via convex cones

We work with the space $\mathcal{H}$ of *hollow symmetric matrices with zero mean off-diagonal entries*, viewed as a subspace of the space $\mathrm{Sym}(P) \coloneqq \{\mathbf{M} \in \mathbb{R}^{P \times P} \mid \mathbf{M} = \mathbf{M}^\top\}$ of symmetric matrices:

$$\mathcal{H} \coloneqq \{\mathbf{M} \in \mathrm{Sym}(P) \mid \operatorname{diag}(\mathbf{M}) = \mathbf{0},\ \mathbf{1}^\top \mathbf{M} \mathbf{1} = 0\}, \tag{159}$$

equipped with the Frobenius inner product $\langle \mathbf{M}, \mathbf{N} \rangle_F \coloneqq \operatorname{tr}(\mathbf{M}^\top \mathbf{N})$ and the corresponding norm $\|\mathbf{M}\|_F^2 = \langle \mathbf{M}, \mathbf{M} \rangle_F$. The orthogonal projection $\Pi_{\mathcal{H}} \colon \mathrm{Sym}(P) \to \mathcal{H}$ is obtained by first removing the diagonal of a symmetric input and then projecting orthogonally to the off-diagonal mean direction $\mathbf{1}\mathbf{1}^\top - \mathbf{I}$, i.e.,

$$\Pi_{\mathcal{H}}(\mathbf{M}) = \mathbf{M} - \operatorname{diag}(\mathbf{M}) - \frac{\langle \mathbf{M}, \mathbf{1}\mathbf{1}^\top - \mathbf{I} \rangle_F}{P(P-1)} (\mathbf{1}\mathbf{1}^\top - \mathbf{I}). \tag{160}$$

Proof of Proposition B.2. The generator set defining $\mathcal{C}_{[k+1]}$ contains the generator set defining $\mathcal{C}_{[k]}$, so $\mathcal{C}_{[k]} \subseteq \mathcal{C}_{[k+1]}$ by monotonicity of the conic hull, establishing the nested chain.

For the strict-inclusion statement, suppose $\Pi_{\mathcal{H}}(\mathbf{u}_{k+1}\mathbf{u}_{k+1}^\top) \notin \mathcal{C}_{[k]}$. Since $\Pi_{\mathcal{H}}(\mathbf{u}_{k+1}\mathbf{u}_{k+1}^\top) \in \mathcal{C}_{[k+1]}$ by definition, the fact that $\mathcal{C}_{[k]} \subsetneq \mathcal{C}_{[k+1]}$ is immediate. Conversely, suppose $\Pi_{\mathcal{H}}(\mathbf{u}_{k+1}\mathbf{u}_{k+1}^\top) \in \mathcal{C}_{[k]}$. Then every nonnegative combination of the generators of $\mathcal{C}_{[k+1]}$ can be rewritten as a nonnegative combination of generators of $\mathcal{C}_{[k]}$ alone, by absorbing the coefficient of $\Pi_{\mathcal{H}}(\mathbf{u}_{k+1}\mathbf{u}_{k+1}^\top)$ into a representation of itself in $\mathcal{C}_{[k]}$. Hence $\mathcal{C}_{[k+1]} \subseteq \mathcal{C}_{[k]}$, and combined with the nested-chain inclusion this gives equality. □

Remark H.8 (Cone relaxation). The cone $\mathcal{C}_{[k]}$ admits arbitrary *nonnegative* coefficients on its generators. The RSMs actually induced by parameterizations $\boldsymbol{\theta} \in \mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ are more constrained: their coefficients on the rank-one terms $\Pi_{\mathcal{H}}(\mathbf{z}_j\mathbf{z}_j^\top)$ and $\Pi_{\mathcal{H}}(\mathbf{u}_i\mathbf{u}_i^\top)$ are *strictly positive* in the positively-homogeneous case and *integer* otherwise (Lemma H.11). The exact achievable set is therefore a proper subset of $\mathcal{C}_{[k]}$, dense in it in the positively-homogeneous case and a discrete subset otherwise.

H.2.3. From cone freedom to ambiguity in similarity scores

The space $\mathcal{H}$ defined in Equation (159), equipped with the Frobenius norm, is isometrically isomorphic to the subspace

$$\mathcal{H}_2 \coloneqq \{\mathbf{x} \in \mathbb{R}^{P(P-1)/2} \mid \mathbf{x}^\top \mathbf{1} = 0\}, \qquad \dim(\mathcal{H}_2) = \frac{P(P-1)}{2} - 1, \tag{161}$$

equipped with the Euclidean norm $\|\mathbf{x}\|^2 \coloneqq \mathbf{x}^\top\mathbf{x}$, via the linear isometry

$$\mathcal{H} \to \mathcal{H}_2, \qquad \mathbf{M} \mapsto \sqrt{2}\,\mathrm{vech}(\mathbf{M}), \tag{162}$$

where vech is the half-vectorization stacking the strict upper triangular entries of $\mathbf{M}$ into a vector.

Proof of Lemma B.3. Let $\mathbf{m}, \mathbf{n} \in \mathbb{R}^{P(P-1)/2}$ denote the vectors of strict upper triangular entries of $\mathbf{M}$ and $\mathbf{N}$, respectively. Pearson correlation between $\mathbf{m}$ and $\mathbf{n}$ equals the Euclidean cosine similarity of their mean-centered versions $\mathbf{m} - \bar{m}\mathbf{1}$ and $\mathbf{n} - \bar{n}\mathbf{1}$, a well-established fact for vectors in $\mathbb{R}^{P(P-1)/2}$. These mean-centered vectors lie in $\mathcal{H}_2$ by construction. Applying the inverse of the isometry established in Equation (162) maps $\mathbf{m} - \bar{m}\mathbf{1}$ and $\mathbf{n} - \bar{n}\mathbf{1}$ to $\Pi_{\mathcal{H}}(\mathbf{M})$ and $\Pi_{\mathcal{H}}(\mathbf{N})$, respectively, and preserves cosine similarity. Hence the Euclidean cosine similarity of $\mathbf{m} - \bar{m}\mathbf{1}$ and $\mathbf{n} - \bar{n}\mathbf{1}$ equals the Frobenius cosine similarity of $\Pi_{\mathcal{H}}(\mathbf{M})$ and $\Pi_{\mathcal{H}}(\mathbf{N})$, giving the claimed identity. □

Proof of Proposition B.4. Throughout this proof, we assume that $\mathcal{C}_{[k]}$ is *pointed*, see Remark H.9.

Let

$$S_{[k]} \coloneqq \{\mathbf{M} \in \mathcal{C}_{[k]} \mid \|\mathbf{M}\|_F = 1\} \tag{163}$$

denote the intersection of $\mathcal{C}_{[k]}$ with the unit sphere in $\mathcal{H}$. Since $\mathcal{C}_{[k]}$ is finitely generated and hence closed, and the unit sphere is compact, $S_{[k]}$ is compact. Pointedness of $\mathcal{C}_{[k]}$ ensures that $S_{[k]}$ is connected.

By Lemma B.3, the map $\mathbf{M} \mapsto \rho(\mathbf{M}, \mathbf{N})$ for $\mathbf{M} \in \mathcal{H} \setminus \{\mathbf{0}\}$ depends only on $\mathbf{M}/\|\mathbf{M}\|_F$, so it suffices to compute its image on $S_{[k]}$. Restricted to $S_{[k]}$, this map is continuous, and its image is therefore a connected compact subset of $[-1, 1]$, hence a closed interval $[\rho_-^{(k)}, \rho_+^{(k)}]$. This image equals $\mathcal{S}_{[k]}(\mathbf{N})$, completing the first claim.

Nesting of the intervals follows immediately from $\mathcal{C}_{[k]} \subseteq \mathcal{C}_{[k+1]}$ (Proposition B.2), since taking the image preserves set inclusion. □

Remark H.9 (Pointedness assumption). Proposition B.4 assumes that the cone $\mathcal{C}_{[k]}$ is pointed, i.e., $\mathcal{C}_{[k]} \cap -\mathcal{C}_{[k]} = \{\mathbf{0}\}$. Pointedness fails precisely when two of the projected generators are negative multiples of each other. In this case, $S_{[k]}$ is disconnected, and $\mathcal{S}_{[k]}(\mathbf{N})$ may consist of finitely many disjoint closed intervals or isolated points rather than a single interval.

H.3. Proofs for Section C

This subsection contains the proofs for the results in Section C. We begin by establishing the notation used throughout, including a unified parameterization that absorbs the positive-scaling and sign-flip symmetries into a single scalar per neuron (Appendix H.3.1). This parameterization yields a clean per-row decomposition of the MRNP and MWNP objectives (Appendix H.3.2), which in turn enables a sequence of partial minimizations. We first eliminate the readout-weight variables in closed form via a constrained quadratic minimization, reducing each per-row objective to a one-dimensional function of an effective weight γ_i (Appendix H.3.3). We then minimize this one-dimensional function over the achievable values of γ_i for irreducible neurons, with the feasible set being either continuous or discrete depending on whether the activation σ is positively homogeneous of degree 1 (Appendix H.3.4). Next, we treat the symmetry-induced features generated by neurons not associated with the irreducible representative, distinguishing activations that exactly attain zero from those that only approach it asymptotically (Appendix H.3.5). Finally, we assemble these per-row optima into the canonical form of the RSM established in Proposition B.1, proving identifiability of the RSM under MRNP and MWNP across all activations of interest (Appendix H.3.6).

H.3.1. Setup

We first establish the notation to be used throughout Appendix H.3. Let

$$\boldsymbol{\theta}^\star = (\mathbf{w}_j^\star, b_j^\star, \mathbf{a}_j^\star)_{j=1}^{N_h^\star} \tag{164}$$

denote the parameter vector of an irreducible representative, and let $\mathbf{H}^\star$ denote its induced hidden activation matrix. Let

$$\boldsymbol{\theta} = (\mathbf{w}_i, b_i, \mathbf{a}_i)_{i=1}^{N_h} \tag{165}$$

denote the parameter vector of a layer in the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ at width $N_h \geq N_h^\star$, and write its induced hidden activation matrix in canonical form

$$\mathbf{H} = \operatorname{diag}(\boldsymbol{\alpha})\mathbf{D}_{\boldsymbol{\nu}} \begin{bmatrix} \mathbf{H}^\star \\ \mathbf{U} \end{bmatrix}, \qquad \boldsymbol{\alpha} \in \mathbb{R}_{\neq 0}^{N_h}, \quad \boldsymbol{\nu} \in \mathbb{N}_{>0}^{N_h^\star + K}, \quad \mathbf{U} \in \mathbb{R}^{K \times P}, \tag{166}$$

according to Proposition 4.2. To account for duplicates, we index the layer's neurons by pairs (i,k) where $i \in \{1, \ldots, N_h^\star + K\}$ selects the unique neuron generating the ith row $\boldsymbol{\psi}_i^\top \in \mathbb{R}^{1\times P}$ of

$$\begin{bmatrix} \mathbf{H}^\star \\ \mathbf{U} \end{bmatrix}, \tag{167}$$

and $k \in \{1, \ldots, \nu_i\}$ picks out one of the ν_i duplicates of that neuron. That is, we index neurons in the hidden layer via the mapping

$$(i,k) \mapsto \nu_{<i} + k, \qquad i = 1, \ldots, N_h^\star + K, \quad k = 1, \ldots, \nu_i, \tag{168}$$

which uses the shorthand notation established in Equation (105). Consequently, the parameters of the (i,k)th neuron are $\mathbf{w}_i$, b_i, and $\mathbf{a}_{i,k}$. Without loss of generality, we assume that the neurons in the overparameterized layer are ordered such that

$$(\mathbf{w}_i, b_i) = (\mathbf{w}_i^\star, b_i^\star), \qquad i = 1, \ldots, N_h^\star. \tag{169}$$

To account for the positive-scaling symmetry (Appendix E.1.2) and the sign-flip symmetry (Appendix E.1.3), we parameterize the (i,k)th neuron as

$$(\alpha_{i,k}\mathbf{w}_i,\ \alpha_{i,k} b_i,\ \alpha_{i,k}^{-1}\mathbf{a}_{i,k}), \qquad i = 1, \ldots, N_h^\star + K, \quad k = 1, \ldots, \nu_i, \tag{170}$$

so that $\alpha_{i,k}^{-1}\mathbf{a}_{i,k}$ is the readout weight of the (i,k)th neuron, where the admissible values of $\alpha_{i,k}$ depend on the activation function σ:

- For positively homogeneous activations of degree 1, $\alpha_{i,k} > 0$.
- For odd activations, $\alpha_{i,k} \in \{-1, +1\}$.
- For activations that are neither positively homogeneous of degree 1 nor odd, $\alpha_{i,k} = 1$.

In all three cases, σ satisfies the scaling property

$$\sigma(\alpha_{i,k} z) = \alpha_{i,k}\, \sigma(z), \qquad z \in \mathbb{R}, \tag{171}$$

for every admissible $\alpha_{i,k}$. When σ is positively homogeneous of degree 1, this is the defining property; when σ is odd, it follows from $\sigma(-z) = -\sigma(z)$; and when σ is neither, $\alpha_{i,k} = 1$ makes it trivial. Consequently, the (i,k)th neuron generates the hidden activity

$$\sigma(\alpha_{i,k}\mathbf{w}_i^\top \mathbf{x} + \alpha_{i,k} b_i) = \alpha_{i,k}\, \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) \tag{172}$$

and contributes

$$\alpha_{i,k}^{-1}\mathbf{a}_{i,k}\, \alpha_{i,k}\, \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) = \mathbf{a}_{i,k}\, \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) \tag{173}$$

to the layer output, leaving the function this neuron computes invariant under $\alpha_{i,k}$. Summing over the ν_i duplicates of the ith unique neuron and requiring that the layer parameterized by $\boldsymbol{\theta}$ compute the same function as its irreducible representative $\boldsymbol{\theta}^\star$ gives, for every $i = 1, \dots, N_h^\star + K$,

$$\sum_{k=1}^{\nu_i} \mathbf{a}_{i,k} = \mathbf{c}_i, \qquad \mathbf{c}_i \coloneqq \begin{cases} \mathbf{a}_i^\star, & i \le N_h^\star, \\ \mathbf{0}, & i > N_h^\star. \end{cases} \tag{174}$$

This constraint involves only the $\mathbf{a}$-coordinates and is independent of the α-coordinates. The feasible region is thus a Cartesian product in the $(\alpha, \mathbf{a})$ parameterization, which allows us to minimize over $\mathbf{a}$ and α iteratively in either order; a fact we exploit in Appendices H.3.3 and H.3.4 below. Finally, the squared norm of the (i,k)th neuron's readout weight $\alpha_{i,k}^{-1}\mathbf{a}_{i,k}$ is

$$\|\alpha_{i,k}^{-1}\mathbf{a}_{i,k}\|^2 = \frac{\|\mathbf{a}_{i,k}\|^2}{\alpha_{i,k}^2}. \tag{175}$$

H.3.2. Row-wise decomposition of the norm-minimizing objectives

We now decompose the MRNP and MWNP objectives into a sum of per-row contributions. Throughout this subsection, we will make repeated use of the fact that the squared Frobenius norm of a matrix coincides with the sum of the squared Euclidean norms of its row, respectively column, vectors. Since the (i,k)th row of $\mathbf{H}$ is given by $\alpha_{i,k}\boldsymbol{\psi}_i^\top$, where $\boldsymbol{\psi}_i^\top = \sigma(\mathbf{w}_i^\top \mathbf{X} + b_i \mathbf{1}^\top)$, the squared Frobenius norm of the hidden activation matrix is

$$\|\mathbf{H}\|_F^2 = \sum_{i=1}^{N_h^\star + K} \|\boldsymbol{\psi}_i\|^2 \sum_{k=1}^{\nu_i} \alpha_{i,k}^2. \tag{176}$$

Similarly, since the readout weight of the (i,k)th neuron is $\alpha_{i,k}^{-1}\mathbf{a}_{i,k}$, the squared Frobenius norm of the readout-weight matrix is

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^{N_h^\star + K} \sum_{k=1}^{\nu_i} \|\alpha_{i,k}^{-1}\mathbf{a}_{i,k}\|^2 = \sum_{i=1}^{N_h^\star + K} \sum_{k=1}^{\nu_i} \frac{\|\mathbf{a}_{i,k}\|^2}{\alpha_{i,k}^2}, \tag{177}$$

where the second identity follows from Equation (175). Thus $\mathcal{J}_{\mathrm{MRNP}} = \|\mathbf{H}\|_F^2 + \|\mathbf{A}\|_F^2$ decomposes as

$$\mathcal{J}_{\mathrm{MRNP}} = \sum_{i=1}^{N_h^\star + K} \left[\|\boldsymbol{\psi}_i\|^2 \sum_{k=1}^{\nu_i} \alpha_{i,k}^2 + \sum_{k=1}^{\nu_i} \frac{\|\mathbf{a}_{i,k}\|^2}{\alpha_{i,k}^2} \right]. \tag{178}$$

The MWNP objective $\mathcal{J}_{\mathrm{MWNP}} = \|\mathbf{W}\|_F^2 + \|\mathbf{b}\|^2 + \|\mathbf{A}\|_F^2$ allows for a similar decomposition. From

$$\|\mathbf{W}\|_F^2 = \sum_{i=1}^{N_h^\star + K} \|\mathbf{w}_i\|^2 \sum_{k=1}^{\nu_i} \alpha_{i,k}^2, \qquad \|\mathbf{b}\|^2 = \sum_{i=1}^{N_h^\star + K} b_i^2 \sum_{k=1}^{\nu_i} \alpha_{i,k}^2, \tag{179}$$

it follows directly that

$$\mathcal{J}_{\mathrm{MWNP}} = \sum_{i=1}^{N_h^\star + K} \left[\left(\|\mathbf{w}_i\|^2 + b_i^2\right) \sum_{k=1}^{\nu_i} \alpha_{i,k}^2 + \sum_{k=1}^{\nu_i} \frac{\|\mathbf{a}_{i,k}\|^2}{\alpha_{i,k}^2} \right]. \tag{180}$$

By writing

$$s_i := \begin{cases} \|\boldsymbol{\psi}_i\|^2 = \|\sigma(\mathbf{w}_i^\top \mathbf{X} + b_i \mathbf{1}^\top)\|^2, & \text{for MRNP} \\ \|\mathbf{w}_i\|^2 + b_i^2, & \text{for MWNP} \end{cases} \tag{181}$$

we can state both the MRNP and MWNP objectives as sums $\sum_{i=1}^{N_h^\star+K} \mathcal{J}^{(i)}$, where

$$\mathcal{J}^{(i)} := s_i \sum_{k=1}^{\nu_i} \alpha_{i,k}^2 + \sum_{k=1}^{\nu_i} \frac{\|\mathbf{a}_{i,k}\|^2}{\alpha_{i,k}^2}. \tag{182}$$

Combining Equation (182) with the function-preserving constraint of Equation (174), the MRNP and MWNP problems take the unified form

$$\min_{\nu_i,\, \alpha_{i,k},\, \mathbf{a}_{i,k}} \sum_{i=1}^{N_h^\star+K} \mathcal{J}^{(i)} \qquad \text{s.t.} \qquad \sum_{k=1}^{\nu_i} \mathbf{a}_{i,k} = \mathbf{c}_i, \quad i = 1, \ldots, N_h^\star + K, \tag{183}$$

with the generating pairs $(\mathbf{w}_i, b_i)$ being *additional* free variables for $i > N_h^\star$, and with the admissible range of $\alpha_{i,k}$ determined by the activation σ as listed in Appendix H.3.1.

H.3.3. Weight-split minimization

The following result minimizes the readout-weight component of Equation (182) for *fixed* duplication ν and scaling parameters $\alpha_1, \ldots, \alpha_\nu$. It holds in all three cases of admissible $\alpha_{i,k}$ enumerated in Appendix H.3.1.

Lemma H.10 (Optimal weight split)**.** *Let $\nu \geq 1$ be an integer, $\alpha_1, \ldots, \alpha_\nu \in \mathbb{R}_{\neq 0}$, and $\mathbf{c} \in \mathbb{R}^{N_o}$. Then the constrained optimization problem*

$$\min_{\mathbf{a}_k \in \mathbb{R}^{N_o}} \sum_{k=1}^{\nu} \frac{\|\mathbf{a}_k\|^2}{\alpha_k^2} \qquad \textit{s.t.} \quad \sum_{k=1}^{\nu} \mathbf{a}_k = \mathbf{c} \tag{184}$$

has the unique minimum

$$\frac{\|\mathbf{c}\|^2}{\sum_{k=1}^{\nu} \alpha_k^2}, \tag{185}$$

attained at

$$\mathbf{a}_k = \frac{\alpha_k^2}{\sum_{m=1}^{\nu} \alpha_m^2} \mathbf{c}, \qquad k = 1, \ldots, \nu. \tag{186}$$

Proof. The Lagrangian

$$\mathcal{L}(\mathbf{a}_1, \ldots, \mathbf{a}_\nu, \boldsymbol{\lambda}) = \sum_{k=1}^{\nu} \frac{\|\mathbf{a}_k\|^2}{\alpha_k^2} - \boldsymbol{\lambda}^\top \left(\sum_{k=1}^{\nu} \mathbf{a}_k - \mathbf{c} \right) \tag{187}$$

has stationarity conditions $2\mathbf{a}_k/\alpha_k^2 = \boldsymbol{\lambda}$ for every k. Hence each $\mathbf{a}_k$ is a scalar multiple of the common vector $\boldsymbol{\lambda}$, namely $\mathbf{a}_k = \frac{1}{2}\alpha_k^2 \boldsymbol{\lambda}$. The constraint $\sum_k \mathbf{a}_k = \mathbf{c}$ then gives $\boldsymbol{\lambda} = 2\mathbf{c}/(\sum_m \alpha_m^2)$, hence $\mathbf{a}_k = \mathbf{c}\, \alpha_k^2/(\sum_m \alpha_m^2)$. Substituting back,

$$\sum_{k=1}^{\nu} \frac{\|\mathbf{a}_k\|^2}{\alpha_k^2} = \|\mathbf{c}\|^2 \sum_{k=1}^{\nu} \frac{\alpha_k^2}{(\sum_m \alpha_m^2)^2} = \frac{\|\mathbf{c}\|^2}{\sum_k \alpha_k^2}. \tag{188}$$

Since each $\alpha_k \neq 0$, the objective is strictly convex in $(\mathbf{a}_1, \ldots, \mathbf{a}_\nu)$, and the feasible set is affine. Therefore the feasible stationary point found above is the unique global minimizer. □

For any fixed duplication ν_i and scaling vector $(\alpha_{i,1}, \ldots, \alpha_{i,\nu_i})^\top$, Lemma H.10 gives the optimal $\mathbf{a}_{i,k}^\star$ in closed form and yields the partially-minimized per-row objective

$$s_i \sum_{k=1}^{\nu_i} \alpha_{i,k}^2 + \frac{\|\mathbf{c}_i\|^2}{\sum_{k=1}^{\nu_i} \alpha_{i,k}^2}. \tag{189}$$

This corresponds to the inner minimization of Equation (183) with $\nu_i, \alpha_{i,k}$, and, for $i > N_h^\star$, the generating pair $(\mathbf{w}_i, b_i)$ held fixed. By introducing the *effective weight*

$$\gamma_i \coloneqq \sum_{k=1}^{\nu_i} \alpha_{i,k}^2, \tag{190}$$

this partially optimized per-row objective can be rewritten as

$$F_i(\gamma_i; \mathbf{w}_i, b_i) \coloneqq \gamma_i \, s_i + \gamma_i^{-1} \|\mathbf{c}_i\|^2, \tag{191}$$

where the neuron parameters $(\mathbf{w}_i, b_i)$ enter through the quantity s_i defined in Equation (181). Minimizing $F_i(\gamma_i; \mathbf{w}_i, b_i)$ gives rise to two distinct regimes.

- For $i \leq N_h^\star$, the parameters $(\mathbf{w}_i, b_i)$ are associated with an "irreducible" hidden neuron producing one of the rows in $\mathbf{H}^\star$, and are hence pinned down by the irreducible representative's parameters $(\mathbf{w}_i^\star, b_i^\star)$. Minimizing Equation (191) then reduces to a one-dimensional minimization objective. This case is addressed in Appendix H.3.4.
- For $i > N_h^\star$, the parameters $(\mathbf{w}_i, b_i)$ are *additional* free parameters that can be tuned to minimize Equation (191). This case is taken care of in Appendix H.3.5.

In both regimes, the optimal $\mathbf{a}_{i,k}^\star$ is recovered from any $(\nu_i^\star, \alpha_{i,k}^\star)$ realizing $\gamma_i^\star$ via Lemma H.10.

H.3.4. One-dimensional minimization over the effective weight

For irreducible rows $i \leq N_h^\star$, the generating pair $(\mathbf{w}_i, b_i) = (\mathbf{w}_i^\star, b_i^\star)$ is pinned down by the irreducible representative, so s_i is fixed and the remaining task is to minimize the one-dimensional function $F_i(\gamma) = \gamma \, s_i + \gamma^{-1} \|\mathbf{c}_i\|^2$ over the achievable values of the effective weight γ. This achievable set depends on whether the activation σ admits the positive-scaling symmetry, as established by the next lemma.

Lemma H.11 (Feasible set of the effective weight)**.** *Considered in isolation from other neurons in the hidden layer, the effective weight γ_i ranges over the set*

$$\Gamma = \begin{cases} (0, \infty), & \text{if } \sigma \text{ is positively homogeneous of degree } 1 \\ \{1, \ldots, N_h - (N_h^\star + K) + 1\}, & \text{otherwise} \end{cases} \tag{192}$$

Proof. If σ is positively homogeneous of degree 1, any $\gamma > 0$ is achieved by taking $\nu_i = 1$ and $\alpha_{i,1} = \sqrt{\gamma}$. Conversely, every choice of positive $\alpha_{i,k}$ gives $\gamma_i > 0$, so $\Gamma = (0, \infty)$.

If σ is not positively homogeneous of degree 1, the admissible values of $\alpha_{i,k}$ all satisfy $\alpha_{i,k}^2 = 1$. Hence,

$$\gamma_i = \sum_{k=1}^{\nu_i} \alpha_{i,k}^2 = \nu_i. \tag{193}$$

The duplication pattern $\boldsymbol{\nu} \in \mathbb{N}_{>0}^{N_h^\star + K}$ has each $\nu_i \geq 1$ and satisfies

$$\sum_{i=1}^{N_h^\star + K} \nu_i = N_h, \tag{194}$$

so that the duplicate count ν_i of the ith neuron is bounded above by $N_h - (N_h^\star + K) + 1$, attained when every other neuron has no duplicates. Thus, $\Gamma = \{1, \ldots, N_h - (N_h^\star + K) + 1\}$ for activations that are not positively homogeneous of degree 1. □

Lemma H.11 discusses individual duplication counts ν_i in isolation. Clearly, for a hidden layer *of a fixed width*, the joint constraint formulated by Equation (194) poses an additional restriction on the set of feasible γ_i that is not taken into account by Lemma H.11.

We now turn to the per-row minimization itself, considering the two regimes (continuous and discrete minimization, depending on σ) in turn.

Lemma H.12 (Continuous minimization). *Let $s > 0$ and $\mathbf{c} \in \mathbb{R}^{N_o}$. If $\mathbf{c} \neq \mathbf{0}$, the function*

$$F\colon (0,\infty) \to (0,\infty), \quad \gamma \mapsto \gamma s + \gamma^{-1}\|\mathbf{c}\|^2 \tag{195}$$

has the unique minimizer $\gamma^\star = \|\mathbf{c}\|/\sqrt{s}$ with $F(\gamma^\star) = 2\sqrt{s}\|\mathbf{c}\|$. For $\mathbf{c} = \mathbf{0}$, $F(\gamma) = \gamma s$ has infimum 0 attained only in the limit $\gamma \to 0$.

Proof. For $\mathbf{c} \neq \mathbf{0}$, differentiating with respect to γ gives $F'(\gamma) = s - \gamma^{-2}\|\mathbf{c}\|^2$, which vanishes uniquely at $\gamma^\star = \|\mathbf{c}\|/\sqrt{s}$. This is indeed a global minimum since $F''(\gamma) = 2\gamma^{-3}\|\mathbf{c}\|^2 > 0$. The minimum value is $F(\gamma^\star) = \gamma^\star s + (\gamma^\star)^{-1}\|\mathbf{c}\|^2 = 2\sqrt{s}\|\mathbf{c}\|$. The $\mathbf{c} = \mathbf{0}$ case is immediate. □

Lemma H.13 (Discrete minimization). *Let $s > 0$ and $\mathbf{c} \in \mathbb{R}^{N_o}$ and write $r := \|\mathbf{c}\|/\sqrt{s}$. If $\mathbf{c} \neq \mathbf{0}$, the function*

$$F\colon \mathbb{N}_{>0} \to (0,\infty), \quad \nu \mapsto \nu s + \nu^{-1}\|\mathbf{c}\|^2 \tag{196}$$

has a unique minimizer except when $r = \sqrt{\nu(\nu+1)}$ for some $\nu \in \mathbb{N}_{>0}$. More precisely, $\nu^\star = \nu$ is the unique minimizer if

$$\sqrt{(\nu-1)\nu} < r < \sqrt{\nu(\nu+1)} \tag{197}$$

for some $\nu \in \mathbb{N}_{>0}$. When $r = \sqrt{\nu(\nu+1)}$, both ν and $\nu+1$ are minimizers. If $\mathbf{c} = \mathbf{0}$, the unique minimizer is $\nu^\star = 1$.

Proof. The forward difference

$$F(\nu+1) - F(\nu) = s - \frac{\|\mathbf{c}\|^2}{\nu(\nu+1)} \tag{198}$$

is strictly increasing in ν, so F is discrete convex on $\mathbb{N}_{>0}$. Hence a point $\nu \geq 2$ is a minimizer precisely when

$$F(\nu) - F(\nu-1) \leq 0 \qquad \text{and} \qquad F(\nu+1) - F(\nu) \geq 0. \tag{199}$$

These inequalities are equivalent to

$$r \geq \sqrt{(\nu-1)\nu} \qquad \text{and} \qquad r \leq \sqrt{\nu(\nu+1)}. \tag{200}$$

For $\nu = 1$, the corresponding condition is simply $F(2) - F(1) \geq 0$, equivalently $r \leq \sqrt{2}$.

If both inequalities are strict, the minimizer is unique. If $r = \sqrt{\nu(\nu+1)}$, then $F(\nu+1) = F(\nu)$, and the two minimizers are ν and $\nu+1$. Given that the forward differences are strictly increasing, this is the only non-unique case.

Finally, if $\mathbf{c} = \mathbf{0}$, then $F(\nu) = \nu s$, which is uniquely minimized at $\nu = 1$. □

This completes the analysis of neurons giving rise to irreducible features, i.e., rows of $\mathbf{H}^\star$. We now turn to the symmetry-induced features, i.e., rows of $\mathbf{U}$, for which the generating pair $(\mathbf{w}_i, b_i)$ is itself a free variable, in Appendix H.3.5.

H.3.5. Symmetry-induced hidden features and their RSM contributions

For neurons generating one of the symmetry-induced hidden features, i.e., rows of $\mathbf{U}$, $i > N_h^\star$, and $\mathbf{c}_i = \mathbf{0}$, Lemma H.10 (applied with $\mathbf{c} = \mathbf{0}$) forces $\mathbf{a}_{i,k}^\star = \mathbf{0}$ for every k, and the per-row objective in Equation (191) collapses to

$$F_i(\gamma_i; \mathbf{w}_i, b_i) = \gamma_i\, s_i. \tag{201}$$

The effective weight γ_i defined in Equation (190) is strictly positive, and s_i is non-negative, so $F_i(\gamma_i; \mathbf{w}_i, b_i) \geq 0$ with equality if and only if $s_i = 0$. Thus, the behavior of these symmetry-induced rows is entirely determined by the ways in which s_i can vanish, which differ between MWNP and MRNP.

MWNP. Since $s_i = \|\mathbf{w}_i\|^2 + b_i^2$, the unique minimizer is the trivial choice

$$(\mathbf{w}_i, b_i) = (\mathbf{0}, 0), \tag{202}$$

attaining $s_i = 0$ for every activation σ. Whether the resulting symmetry-induced hidden feature vanishes altogether or is merely constant depends on the activation σ. If $\sigma(0) = 0$, then $\boldsymbol{\psi}_i = \mathbf{0}^\top$. Otherwise, $\boldsymbol{\psi}_i = \sigma(0)\mathbf{1}^\top$ is a constant nonzero row.

MRNP. Since $s_i = \|\sigma(\mathbf{w}_i^\top \mathbf{X} + b_i \mathbf{1}^\top)\|^2$, attaining $s_i = 0$ requires the symmetry-induced hidden feature itself to vanish:

$$\sigma(\mathbf{w}_i^\top \mathbf{X} + b_i \mathbf{1}^\top) = \mathbf{0}^\top. \tag{203}$$

Setting $\mathbf{w}_i = \mathbf{0}$ reduces the pre-activation to the constant row $b_i \mathbf{1}^\top$, and Equation (203) becomes $\sigma(b_i) = 0$, i.e., $b_i \in \sigma^{-1}(\{0\})$. The two cases addressed by the next two lemmata correspond to whether this preimage is empty.

Lemma H.14 (Exact vanishing of symmetry-induced features)**.** *Let σ be an activation function with $\sigma^{-1}(\{0\}) \neq \emptyset$, and pick any $b^\star \in \sigma^{-1}(\{0\})$. Then $(\mathbf{w}, b) = (\mathbf{0}, b^\star)$ realizes*

$$\sigma(\mathbf{w}^\top \mathbf{X} + b \mathbf{1}^\top) = \mathbf{0}^\top. \tag{204}$$

Proof. Setting $(\mathbf{w}, b) = (\mathbf{0}, b^\star)$ gives the pre-activation $\mathbf{0}^\top \mathbf{X} + b^\star \mathbf{1}^\top = b^\star \mathbf{1}^\top$, and applying σ elementwise yields $\sigma(b^\star)\mathbf{1}^\top = \mathbf{0}^\top$ by the choice of $b^\star$. □

For activations covered by Lemma H.14, all activations in Table F.1 except log-sigmoid, sigmoid, softplus, and squareplus, the MRNP global minimum $F_i = 0$ is *attained* at $(\mathbf{w}_i, b_i) = (\mathbf{0}, b^\star)$ for any $b^\star \in \sigma^{-1}(\{0\})$, with $\boldsymbol{\psi}_i = \mathbf{0}^\top$ vanishing outright.

Lemma H.15 (Asymptotic vanishing of symmetry-induced features)**.** *Let σ be an activation function with $\sigma^{-1}(\{0\}) = \emptyset$ that vanishes at infinity, in the sense that $\lim_{x\to+\infty} \sigma(x) = 0$ or $\lim_{x\to-\infty} \sigma(x) = 0$. Then, there exists a sequence $(\mathbf{w}_t, b_t)_{t\in\mathbb{N}}$ with $\mathbf{w}_t = \mathbf{0}$ and $|b_t| \to \infty$ such that*

$$\sigma(\mathbf{w}_t^\top \mathbf{X} + b_t \mathbf{1}^\top) = \sigma(b_t)\mathbf{1}^\top \xrightarrow{t\to\infty} \mathbf{0}^\top. \tag{205}$$

Proof. Take $\mathbf{w}_t = \mathbf{0}$ and let b_t tend to whichever of $\pm\infty$ realizes $\sigma(b_t) \to 0$. Substituting into the pre-activation gives $\mathbf{0}^\top \mathbf{X} + b_t \mathbf{1}^\top = b_t \mathbf{1}^\top$, and applying σ elementwise yields $\sigma(b_t)\mathbf{1}^\top$, which converges to $\mathbf{0}^\top$ by choice of b_t. □

For activations covered by Lemma H.15, log-sigmoid, sigmoid, softplus, and squareplus from Table F.1, the MRNP global minimum $F_i = 0$ is only *approached* along the sequence $(\mathbf{0}, b_t)$ with $|b_t| \to \infty$, never attained at any finite configuration. Each element of the sequence yields a constant symmetry-induced feature $\sigma(b_t)\mathbf{1}^\top$ whose norm vanishes only in the limit.

H.3.6. Identifiability of the RSM under MRNP and MWNP

We now assemble the per-row optima from the preceding subsections to establish identifiability of the RSM under MRNP and MWNP, treating the positively homogeneous case (Proposition C.3 of the main text) first. Thereafter, we present four additional identifiability results that do not assume the activation to be positively homogeneous of degree 1.

Positively homogeneous case. We now prove the main-text proposition.

Proof of Proposition C.3. By Proposition B.1, every parameterization in the orbit yields an RSM of the form

$$\mathrm{RSM} = \sum_{j=1}^{N_h^\star} \gamma_j \, \mathbf{z}_j \mathbf{z}_j^\top + \sum_{k=1}^{K} \gamma_{N_h^\star + k} \, \mathbf{u}_k \mathbf{u}_k^\top, \tag{206}$$

where $\mathbf{z}_j^\top$ is the jth row of $\mathbf{H}^\star$, $\mathbf{u}_k^\top$ is the kth row of $\mathbf{U}$, and $\boldsymbol{\gamma} = \mathbf{D}_{\boldsymbol{\nu}}^\top \boldsymbol{\alpha}^2$. By the per-row decomposition of Equation (182) together with the function-preserving constraint of Equation (174), both $\mathcal{J}_{\mathrm{MRNP}}$ and $\mathcal{J}_{\mathrm{MWNP}}$ decouple across i. Minimizing the full objective therefore reduces to minimizing each per-row term independently.

For irreducible rows $j \leq N_h^\star$, the assumption $s_j > 0$ together with $\mathbf{c}_j = \mathbf{a}_j^\star \neq \mathbf{0}$, which follows from irreducibility of $\boldsymbol{\theta}^\star$, puts us in the non-degenerate regime of Lemma H.12, which gives the unique minimizer $\gamma_j^\star = \|\mathbf{a}_j^\star\|/\sqrt{s_j}$.

For symmetry-induced rows $i = N_h^\star + k$ with $k = 1, \ldots, K$, positive homogeneity of degree 1 forces $\sigma(0) = 0$, so $0 \in \sigma^{-1}(\{0\})$ and Lemma H.14 applies with $b^\star = 0$. Both the MRNP and MWNP optima are attained at $(\mathbf{w}_i, b_i) = (\mathbf{0}, 0)$, yielding $\mathbf{u}_k = \mathbf{0}^\top$ outright. The symmetry-induced contributions to the RSM thus vanish:

$$\gamma_{N_h^\star + k} \, \mathbf{u}_k \mathbf{u}_k^\top = \mathbf{0}, \qquad k = 1, \ldots, K, \tag{207}$$

regardless of the (free) choice of $\gamma_{N_h^\star + k}$. Substituting into Equation (206) yields Equation (18). □

For the remainder of this subsection, we consider a more general case, dropping the assumption that the activation σ is positively homogeneous of degree 1. In the discrete regime that results, joint feasibility of the per-row optima imposes a width condition that we now formalize.

Assumption H.16 (Ample width)**.** The overparameterized layer's width N_h satisfies

$$N_h \geq \sum_{j=1}^{N_h^\star} \gamma_j^\star + K, \tag{208}$$

where $\gamma_j^\star$ is the per-row effective-weight optimum from Lemma H.13, and K is the number of symmetry-induced features, i.e., the number of rows of $\mathbf{U}$.

Assumption H.16 ensures that the per-row optima identified in Appendices H.3.4 and H.3.5 are jointly realizable: each irreducible neuron can be duplicated $\gamma_j^\star$ times and each symmetry-induced feature can occupy a single additional neuron. The setting in which Assumption H.16 fails is briefly discussed in Remark H.23.

MRNP with *exactly* vanishing symmetry-induced features. For activations not satisfying positive homogeneity but admitting a σ-zero (i.e., $\sigma^{-1}(\{0\}) \neq \emptyset$), the symmetry-induced features still vanish exactly via Lemma H.14, and the only change relative to the positively-homogeneous case is that the irreducible-row optimum is now discrete.

Proposition H.17 (MRNP for activations with $\sigma^{-1}(\{0\}) \neq \emptyset$)**.** *Let $\boldsymbol{\theta}^\star$ be the parameter vector of an irreducible hidden layer with activation σ that is not positively homogeneous of degree 1 but satisfies $\sigma^{-1}(\{0\}) \neq \emptyset$, and let $\mathbf{H}^\star$, $\mathbf{z}_j^\top$, s_j be defined as in Proposition C.3. Assume that $s_j > 0$ for all $j = 1, \ldots, N_h^\star$. Under Assumption H.16, every MRNP in the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ has the same RSM, namely*

$$\mathrm{RSM} = \sum_{j=1}^{N_h^\star} \gamma_j^\star \, \mathbf{z}_j \mathbf{z}_j^\top, \qquad \gamma_j^\star \in \arg\min_{\nu \in \mathbb{N}_{>0}} \left(\nu \, s_j + \nu^{-1} \|\mathbf{a}_j^\star\|^2 \right), \tag{209}$$

where $\gamma_j^\star$ is unique except when $\|\mathbf{a}_j^\star\| / \sqrt{s_j} = \sqrt{\nu(\nu+1)}$ for some $\nu \in \mathbb{N}_{>0}$, in which case $\gamma_j^\star \in \{\nu, \nu+1\}$.

Proof. The argument follows the proof of Proposition C.3 with two changes. First, by Lemma H.11, the effective weight γ_j ranges over $\mathbb{N}_{>0}$ rather than $(0, \infty)$, so the irreducible-row minimizer is given by Lemma H.13 instead of Lemma H.12. Second, although σ is not positively homogeneous, $\sigma^{-1}(\{0\}) \neq \emptyset$ permits invoking Lemma H.14 with any $b^\star \in \sigma^{-1}(\{0\})$, yielding $\mathbf{u}_k = \mathbf{0}^\top$ outright and hence vanishing symmetry-induced contributions to the RSM. □

MWNP with *exactly* vanishing symmetry-induced features. For MWNP under the same activation hypotheses, the symmetry-induced rows are forced to take the form $\sigma(0)\mathbf{1}^\top$ rather than vanishing exactly, so identifiability holds only after projection onto $\mathcal{H}$ unless $\sigma(0) = 0$.

Proposition H.18 (MWNP for activations with $\sigma^{-1}(\{0\}) \neq \emptyset$)**.** *Let $\boldsymbol{\theta}^\star$ be the parameter vector of an irreducible hidden layer with activation σ that is not positively homogeneous of degree 1 but satisfies $\sigma^{-1}(\{0\}) \neq \emptyset$, and let $\mathbf{H}^\star$, $\mathbf{z}_j^\top$, s_j be defined as in Proposition C.3. Under Assumption H.16, every MWNP in the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ has the same projected RSM, namely*

$$\Pi_{\mathcal{H}}(\mathrm{RSM}) = \sum_{j=1}^{N_h^\star} \gamma_j^\star \, \Pi_{\mathcal{H}}(\mathbf{z}_j \mathbf{z}_j^\top), \qquad \gamma_j^\star \in \arg\min_{\nu \in \mathbb{N}_{>0}} \left(\nu \, s_j + \nu^{-1} \|\mathbf{a}_j^\star\|^2 \right), \tag{210}$$

with $\gamma_j^\star$ unique up to the same discrete tie-breaking exception as in Proposition H.17.

Proof. The irreducible-row analysis is identical to that in the proof of Proposition H.17. For symmetry-induced rows, the MWNP optimum is attained at $(\mathbf{w}_i, b_i) = (\mathbf{0}, 0)$ per the analysis in Appendix H.3.5, yielding the constant row $\mathbf{u}_k = \sigma(0)\mathbf{1}^\top$. The corresponding RSM contribution is $\gamma_{N_h^\star + k} \, \sigma(0)^2 \, \mathbf{1}\mathbf{1}^\top$, which lies in the span of $\mathbf{1}\mathbf{1}^\top$ and is therefore in the orthogonal complement of $\mathcal{H}$. Hence,

$$\Pi_{\mathcal{H}}\left(\gamma_{N_h^\star + k} \, \mathbf{u}_k \mathbf{u}_k^\top \right) = \mathbf{0}, \qquad k = 1, \ldots, K, \tag{211}$$

and applying $\Pi_{\mathcal{H}}$ to Equation (206) yields Equation (210). □

When $\sigma(0) = 0$, the constant row $\sigma(0)\mathbf{1}^\top$ in the proof above is the zero row, so the symmetry-induced contributions vanish altogether and the projection $\Pi_{\mathcal{H}}$ in Equation (210) can be dropped, yielding uniqueness of the RSM prior to projecting onto $\mathcal{H}$.

MRNP with *asymptotically* vanishing symmetry-induced features. When $\sigma^{-1}(\{0\}) = \emptyset$, the symmetry-induced features can no longer be made to vanish at any finite parameter configuration, and the MRNP infimum is approached only along a minimizing sequence with $|b_t| \to \infty$.

Proposition H.19 (MRNP for activations with $\sigma^{-1}(\{0\}) = \emptyset$ vanishing at infinity)**.** *Let $\boldsymbol{\theta}^\star$ be the parameter vector of an irreducible hidden layer with activation σ that is not positively homogeneous of degree 1, satisfies $\sigma^{-1}(\{0\}) = \emptyset$, and vanishes at infinity in the sense of Lemma H.15. Let $\mathbf{H}^\star$, $\mathbf{z}_j^\top$, s_j be defined as in Proposition C.3, and assume $s_j > 0$ for all $j = 1, \dots, N_h^\star$. Under Assumption H.16, the MRNP infimum in the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ is approached along a minimizing sequence whose RSM converges to a common limit satisfying*

$$\Pi_{\mathcal{H}}(\mathrm{RSM}) = \sum_{j=1}^{N_h^\star} \gamma_j^\star \, \Pi_{\mathcal{H}}(\mathbf{z}_j \mathbf{z}_j^\top), \qquad \gamma_j^\star \in \operatorname*{arg\,min}_{\nu \in \mathbb{N}_{>0}} \left(\nu \, s_j + \nu^{-1} \|\mathbf{a}_j^\star\|^2 \right), \tag{212}$$

with $\gamma_j^\star$ unique up to the same discrete tie-breaking exception as in Proposition H.17.

Proof. The irreducible-row analysis is identical to that in the proof of Proposition H.17. For symmetry-induced rows, $\sigma^{-1}(\{0\}) = \emptyset$ rules out Lemma H.14, and we instead invoke Lemma H.15: the MRNP infimum is approached along the sequence $(\mathbf{0}, b_t)$ with $|b_t| \to \infty$, yielding constant rows $\mathbf{u}_k^{(t)} = \sigma(b_t)\mathbf{1}^\top$. The corresponding RSM contribution along the sequence is $\gamma_{N_h^\star + k} \, \sigma(b_t)^2 \, \mathbf{1}\mathbf{1}^\top$, which lies in the span of $\mathbf{1}\mathbf{1}^\top$ and is therefore in the orthogonal complement of $\mathcal{H}$. Applying $\Pi_{\mathcal{H}}$ to Equation (206) along the sequence yields Equation (212). □

MWNP with *asymptotically* vanishing symmetry-induced features. The MWNP optimum, in contrast, is still attained at $(\mathbf{w}_i, b_i) = (\mathbf{0}, 0)$ and yields the constant row $\sigma(0)\mathbf{1}^\top$, so identifiability holds after projection onto $\mathcal{H}$ for the same reason as in Proposition H.18.

Proposition H.20 (MWNP for activations with $\sigma^{-1}(\{0\}) = \emptyset$ vanishing at infinity)**.** *Let $\boldsymbol{\theta}^\star$ be the parameter vector of an irreducible hidden layer with activation σ that is not positively homogeneous of degree 1, satisfies $\sigma^{-1}(\{0\}) = \emptyset$, and vanishes at infinity in the sense of Lemma H.15. Let $\mathbf{H}^\star$, $\mathbf{z}_j^\top$, s_j be defined as in Proposition C.3. Under Assumption H.16, every MWNP in the orbit $\mathcal{O}_{N_h}(\boldsymbol{\theta}^\star)$ has the same projected RSM, namely*

$$\Pi_{\mathcal{H}}(\mathrm{RSM}) = \sum_{j=1}^{N_h^\star} \gamma_j^\star \, \Pi_{\mathcal{H}}(\mathbf{z}_j \mathbf{z}_j^\top), \qquad \gamma_j^\star \in \operatorname*{arg\,min}_{\nu \in \mathbb{N}_{>0}} \left(\nu \, s_j + \nu^{-1} \|\mathbf{a}_j^\star\|^2 \right), \tag{213}$$

with $\gamma_j^\star$ unique up to the same discrete tie-breaking exception as in Proposition H.17.

Proof. Mechanically identical to the proof of Proposition H.18: the MWNP optimum is attained at $(\mathbf{w}_i, b_i) = (\mathbf{0}, 0)$, yielding the constant row $\mathbf{u}_k = \sigma(0)\mathbf{1}^\top$, and the corresponding RSM contribution lies in the orthogonal complement of $\mathcal{H}$. The hypothesis $\sigma^{-1}(\{0\}) = \emptyset$ enters only insofar as it implies $\sigma(0) \neq 0$, so the constant row is genuinely nonzero and the projection cannot be dropped (in contrast with the strengthening following Proposition H.18). □

Discussion. The five propositions above establish that MRNP and MWNP select an essentially unique RSM across all activations in Table F.1, with three caveats addressed by the following three remarks.

Remark H.21 (Almost-everywhere uniqueness in the discrete regime). For any fixed irreducible representative, Propositions H.17 to H.20 provide a unique $\gamma_j^\star$ unless $\|\mathbf{a}_j^\star\|/\sqrt{s_j} = \sqrt{\nu(\nu+1)}$ for some $\nu \in \mathbb{N}_{>0}$, in which case both ν and $\nu + 1$ are minimizers. Allowing the irreducible parameters to vary, the ratio $\|\mathbf{a}_j^\star\|/\sqrt{s_j}$ is a smooth function of the parameters wherever $s_j > 0$, and the set on which it equals $\sqrt{\nu(\nu+1)}$ for some $\nu \in \mathbb{N}_{>0}$ is a countable union of smooth hypersurfaces in this region. Uniqueness therefore holds for almost every irreducible representative in the Lebesgue sense.

Remark H.22 (Forced duplication in the discrete regime). The discrete regime exhibits a qualitative behavior absent in the positively homogeneous case: the optimum forces the duplication of the hidden neuron generating $\mathbf{z}_j$ whenever $\|\mathbf{a}_j^\star\|/\sqrt{s_j} > \sqrt{2}$. A neuron with a large readout weight relative to $\sqrt{s_j}$ is thus "diluted" across multiple copies, each

carrying a fraction of the original readout weight. In contrast, the positively homogeneous case admits a continuous scaling that makes duplication a pure gauge freedom: any choice of $\nu_j \geq 1$ with $\sum_k \alpha_{j,k}^2 = \gamma_j^\star$ realizes the same effective weight $\gamma_j^\star$, and hence the same RSM.

Remark H.23 (Beyond the ample-width regime). When Assumption H.16 fails, MRNP and MWNP still exist but solve a budget-constrained variant in which the per-row one-dimensional optima are no longer jointly realizable. The corresponding analysis is beyond the scope of this work.