LAYERWISE AND EXOTIC EQUIVARIANCE CONSTRAINTS IN SHALLOW NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We prove the existence and uniqueness of a canonical form for shallow neural networks with activations that are uncoupling, meaning that the activations take pre-activations to linearly independent functions. Prior work shows that ReLU, Tanh and several other activations are uncoupling. We generalize this to also include modern activations such as GELU, SiLU and ELU. Using the canonical form, we characterize all possible equivariant shallow networks with uncoupling activations in terms of equivariance constraints on the weights and biases. Some of the constraints define classical layerwise equivariant neural networks, but some are exotic variants that are not layerwise equivariant, instead having inter-layer constraints. Our results generalize prior work by Agrawal & Ostrowski (2022), who characterized single output invariant shallow ReLU networks. Furthermore, we show that layerwise networks and exotic networks can have qualitatively different optimization behaviour and test the practical impact of different choices of equivariant structures in the MLP blocks of equivariant vision transformers.

Symmetries can be a useful inductive bias for many learning tasks. One example is equivariance to isometries in models for 3D molecular generation (Irwin et al., 2025). The research field concerned with exploiting this inductive bias has become known as *geometric deep learning* (Bronstein et al., 2017; 2021), with a line of research going back to Wood & Shawe-Taylor (1996) studying so-called *group equivariant* neural networks.

Definition 0.1. A *representation* of a group G on a vector space V is a group homomorphism $\rho_V : G \to \mathrm{GL}(V)$ and V is known as the *representation space*. A map $f : U \to V$ between two representation spaces is called *equivariant* if

$$f(\rho_U(g)u) = \rho_V(g)f(u), \quad \text{for all } g \in G, u \in U. \tag{1}$$

If $\rho_V(g) = \mathrm{id}$, then we say f is *invariant*.

It is easy to see that compositions of equivariant maps are again equivariant. This is the basis for one of the main strategies for constructing equivariant deep neural networks, namely to build equivariance directly into the model *layerwise* by using equivariant activations and restricting the linear layers to an equivariant subspace (Wood & Shawe-Taylor, 1996; Cohen & Welling, 2016). Recently, Shahverdi et al. (2026a) studied in what cases equivariant neural networks can be written on such a layerwise equivariant form. They found a sufficient condition in *parameter identifiability*, meaning when a neural network's weights and biases are determined by its output. Shahverdi et al. (2026a) showed that every layerwise identifiable equivariant neural network is layerwise equivariant. However, they also observed that certain activations σ, such as ReLU and GELU, satisfy

$$\sigma(-t) = \sigma(t) - t, \tag{2}$$

which hinders layerwise identifiability since certain parameters can "pass through" the activation via the linear term. This symmetry of the activation allows for *exotic* equivariant networks.

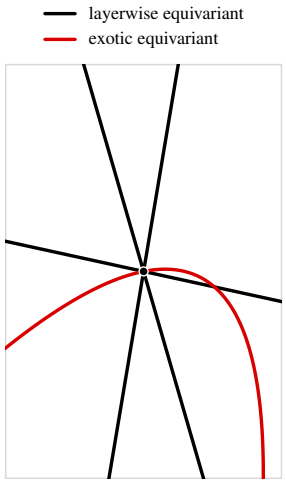

Figure 1: Qualitative visualization of the parameter space of a shallow neural network. We classify all equivariant networks into parameters that satisfy layerwise equivariance constraints (yielding linear parameter subspaces) or exotic equivariance constraints with inter-layer dependencies (not linear parameter subspaces in general).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

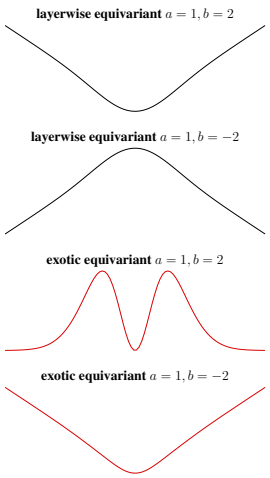

Figure 2: Qualitative plot of the functions parametrized by (4) (black) and (5) (red) for different parameter values. See Appendix D.1 for more details.

Example 0.2 (Exotic GELU network). Let the group $G = \{1, -1\}$ act on two copies of $\mathbb{R}$ by $\rho_{\text{in}}(-1) = -1$ and $\rho_{\text{out}}(-1) = 1$, so that equivariant functions are even functions on $\mathbb{R}$. Let $\sigma = \text{GELU}$ and consider a shallow neural network f_θ, parametrized by $\theta = (W \in \mathbb{R}^{2\times 1}, A \in \mathbb{R}^{1\times 2})$: $f_\theta(x) = A\sigma(Wx)$. Using the symmetry (2),

$$A\sigma(-Wx) = A\sigma(Wx) - AWx, \tag{3}$$

so that f_θ is even if and only if $AW = 0$. Choosing

$$A = (b \quad b)\,,\ W = (a \quad -a)^T\,,\ a, b \in \mathbb{R} \setminus \{0\}, \tag{4}$$

is the standard layerwise permutation equivariant parametrization. Indeed, the only choice of representation on the intermediate layer that commutes with σ is a permutation. However,

$$A = (-b \quad a)\,,\ W = (a \quad b)^T\,,\ a, b \in \mathbb{R} \setminus \{0\}, \tag{5}$$

yields an even function, but requires $a = -b$ for the latent representation to be a permutation. For $a \neq -b$, there is no representation on the latent space with both the linear layers and the activation equivariant. Further, there is a dependency between the parameters in the first and second layer, but no dependency within each layer. We call (5) an exotic parameterization. We plot the two equivariant variants (4) and (5) in Figure 2, showing their qualitative behaviour. By Prop. 1.24, they parameterize different functions in general.

The existence of exotic equivariant networks such as the one in (5) naturally leads to the question of what all possible equivariant networks are. For ReLU activations, this was studied by Agrawal & Ostrowski (2022), who classified all single-output shallow *invariant* ReLU networks. Their classification goes via identifiability in a similar sense to (Shahverdi et al., 2026a). However, the identifiability in (Agrawal & Ostrowski, 2022) is not layerwise, but rather via the existence of a canonical form for an entire shallow network, with possible dependency between the two layers. The dependence is caused by the symmetry (2). In this paper, we generalize the results of Agrawal & Ostrowski (2022) to multi-dimensional output, full equivariance and all activations in Table 1. We will see that many different equivariant weight-tying constraints are possible, some of them layerwise and some exotic. A qualitative picture of the parameter space of a shallow network is shown in Figure 1. For deep neural networks, we know of no results that give a general characterization of equivariant networks. However, Agrawal & Ostrowski (2023) constructed an exotic deep ReLU architecture.

There is a long line of research that studies identifiability of both *shallow* and *deep neural networks*, covering classical bounded activations such as Tanh and Sigmoid (Sussmann, 1992; Albertini et al., 1993; Kůrková & Kainen, 1994; Fefferman, 1994), with the more recent Vlačić & Bölcskei (2021) proving identifiability for a general class of Tanh-like activations. Modern unbounded activations such as ReLU and GELU are less common in the identifiability literature. However, some identifiability results exist for ReLU networks (Grigsby et al., 2023; Bona-Pellissier et al., 2023; Gegenfurtner et al., 2026; Zhang & Bölcskei, 2026). There is also a line of work dealing with identifiability for networks with polynomial activations (Kileel et al., 2019; Finkel et al., 2025; Usevich et al., 2025; Shahverdi et al., 2026b; Kohn et al., 2026). Works addressing identifiability do so up to symmetries such as neuron permutations, since these symmetries prohibit identification. In fact, one of the main concerns of the literature is to find a set of operations modulo which identifiablity is possible. Other prior work studies identifiability empirically. Most relevant to our case, Martinelli et al. (2023) identified the symmetry (2) of "even + linear" and also "odd + constant" activations, and Bökman & Kahl (2023) used identifiability as a motivation to look for equivariance in unconstrained networks.

We restrict our attention to *shallow neural networks*, which are of high practical importance since they are used as components in transformers and other modern architectures (Vaswani et al., 2017).

Definition 0.3. A *shallow neural network* from $\mathbb{R}^m$ to $\mathbb{R}^p$ with activation $\sigma : \mathbb{R} \to \mathbb{R}$ and width n, is parameterized by $W \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^n, A \in \mathbb{R}^{p\times n}, C \in \mathbb{R}^{p\times m}, d \in \mathbb{R}^p$ and defines the function

$$x \mapsto A\sigma(Wx + b) + Cx + d, \tag{6}$$

where σ is applied coordinate-wise.

Table 1: Intrinsic symmetry groups $\Gamma_\sigma \leq \mathrm{Aff}_1 = \{(\alpha, \beta) \in \mathbb{R}^2 \mid \alpha \neq 0\}$, characters χ_σ and residuals r_σ, s_σ of common activation functions. We have $\sigma(\alpha t+\beta) = \chi_\sigma(\alpha, \beta)\sigma(t)+r_\sigma(\alpha, \beta)t+s_\sigma(\alpha, \beta)$ for all $(\alpha, \beta) \in \Gamma_\sigma$. We also list constants p_σ, q_σ such that $\tilde{\sigma}(t) = \sigma(t) + p_\sigma t + q_\sigma$ is centered, i.e. has $r_{\tilde{\sigma}} = s_{\tilde{\sigma}} = 0$.

σ	Γ_σ	$\chi_\sigma(\alpha, \beta)$	$r_\sigma(\alpha, \beta)$	$s_\sigma(\alpha, \beta)$	p_σ	q_σ
Tanh	$\alpha = \pm 1, \beta = 0$	α	0	0	0	0
Sigmoid	$\alpha = \pm 1, \beta = 0$	α	0	1 if $\alpha = -1$, else 0	0	$-1/2$
GELU, SiLU, Softplus	$\alpha = \pm 1, \beta = 0$	1	-1 if $\alpha = -1$, else 0	0	$-1/2$	0
ReLU	$\beta = 0$	$\lvert\alpha\rvert$	α if $\alpha < 0$, else 0	0	$-1/2$	0
Exp	$\alpha = 1$	e^β	0	0	0	0
ELU, SELU	$\alpha = 1, \beta = 0$	1	0	0	0	0

The residual term Cx in Definition 0.3 will be necessary to handle networks with activations whose intrinsic symmetries have a non-trivial linear term (r_σ in Table 1).

The broad route of this paper goes as follows. First, we define Γ_σ, the *group of intrinsic symmetries* (such as (2)) of an activation σ. For instance, $\Gamma_{\mathrm{GELU}} \cong \{1, -1\}$. We define a canonical form for shallow networks, which removes neurons that do not contribute to the network output, normalizes the neurons with respect to intrinsic symmetries Γ_σ and fixes a permutation of the neurons. This canonical form of a network is unique whenever the activation of the network is *uncoupling*, which includes every activation in Table 1. Further, we show that *almost canonical* networks, networks that have no neurons related by Γ_σ and all neurons contributing to the output, are related to their canonical form by a *gauge transformation* of their parameters. It then follows that two almost canonical networks that define the same function are related by a unique gauge. The gauges form the group $\Gamma_\sigma \wr S_n$, which should be thought of as the intrinsic symmetry of networks of width n. For GELU, $\Gamma_\sigma \wr S_n$ consists of signed permutation matrices. For almost canonical networks equivariant under a group G, we employ the adjunction trick by Marchetti et al. (2024); Shahverdi et al. (2026a) to obtain weight-tying conditions determined by a homomorphism $\Psi : G \to \Gamma_\sigma \wr S_n$, and we show that different Ψ yield different realized functions. When the image of Ψ consists of only elements that commute with σ, we get a standard layerwise equivariant network, otherwise we say it is exotic.

Furthermore, we will see that the optimization dynamics of exotic equivariant networks can differ qualitatively from layerwise equivariant networks. In particular, for orthogonal representations, the space of layerwise equivariant parameters is known to be invariant under gradient flow with full group orbit data augmentation (Nordenfors et al., 2025), but we will see that the equivariant parameters in the exotic net in (5) are not. Finally, we train equivariant vision transformers whose MLP blocks are shallow neural networks with several different equivariant parametrizations, finding that the choice impacts the performance and we point out promising directions for future work.

1 The Structure of Shallow Neural Networks

The goal of this section is to develop a generalization of the ReLU-canonical form by Agrawal & Ostrowski (2022) and prove its existence and uniqueness for a broad class of activations. We will then use the canonical form to find weight tying conditions for shallow equivariant networks in Section 2. Proofs are deferred to Appendix B, but for substantial proofs we sketch the main ideas. In the supplementary material, we provide Lean formalizations of all results.

1.1 Intrinsic Symmetries of Activation Functions

We proceed by defining the *intrinsic symmetry* of an activation, which is used to remove "obvious" symmetries that would otherwise obstruct the definition of the canonical form. We also define the notion of an *uncoupling* activation, which is sufficient to prove uniqueness of the canonical form.

Let Aff_1 denote the group of invertible affine transformations of $\mathbb{R}$. Elements of Aff_1 are maps $t \mapsto \alpha t + \beta$, so we can identify the elements with tuples $(\alpha, \beta) \in \mathbb{R}^2$, $\alpha \neq 0$, under the group

composition law $(\alpha_2, \beta_2)(\alpha_1, \beta_1) = (\alpha_2\alpha_1, \alpha_2\beta_1 + \beta_2)$. There is a natural right action of Aff_1 on the vector space of functions $\sigma : \mathbb{R} \to \mathbb{R}$ given by precomposition:

$$[\sigma \cdot (\alpha, \beta)](t) = \sigma(\alpha t + \beta). \tag{7}$$

Definition 1.1. For $\sigma : \mathbb{R} \to \mathbb{R}$ define $\Gamma_\sigma \leq \mathrm{Aff}_1$ to be the stabilizer of $\mathrm{span}\{1, t, \sigma\}$ under the action (7). That is, Γ_σ is the subset of Aff_1 such that for $(\alpha, \beta) \in \Gamma_\sigma$, there exist γ, δ, ϵ with

$$\sigma(\alpha t + \beta) = \gamma\sigma(t) + \delta t + \epsilon. \tag{8}$$

We call Γ_σ the *intrinsic symmetry group* of σ.

An example of an intrinsic symmetry was given in (2): $\sigma(-t) = \sigma(t) - t$. We note that analogous definitions to Definition 1.1 show up in prior work, but these lack the linear term δt. For example, the definition of *self-affine* in Kůrková & Kainen (1994) and Definition 1(i) in Vlačić & Bölcskei (2021). The authors of the latter note that it would be interesting to study symmetries with $\delta \neq 0$.

Proposition 1.2. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-affine function. Then γ, δ, ϵ in* (8) *are unique and $\gamma \neq 0$. We can therefore define maps $\chi_\sigma : \Gamma_\sigma \to \mathbb{R} \setminus \{0\}$ and $r_\sigma : \Gamma_\sigma \to \mathbb{R}$, $s_\sigma : \Gamma_\sigma \to \mathbb{R}$ as $\chi_\sigma(\alpha, \beta) = \gamma$ and $r_\sigma(\alpha, \beta) = \delta, s_\sigma(\alpha, \beta) = \epsilon$ via* (8)*. χ_σ is called the* intrinsic character *of σ, as it is a character of Γ_σ (i.e. χ_σ is a group homomorphism). r_σ, s_σ are called the* intrinsic linear and constant residual *of σ respectively, they are not group homomorphsims in general.*

In Table 1 we list the intrinsic symmetries, characters and residuals of common activation functions.

Definition 1.3. A *pre-activation* is parameterized by a pair (w, b) with $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$. The pre-activation is the function $x \mapsto w \cdot x + b$. A pre-activation is called *degenerate* if $w = 0$. We define an action of Aff_1 on pre-activations via

$$(\alpha, \beta) \cdot (w, b) = (\alpha w, \alpha b + \beta). \tag{9}$$

Given $\Gamma \leq \mathrm{Aff}_1$, two pre-activations (w, b) and (w', b') are called Γ-*separated* if they are in different Γ-orbits under (9).

Example 1.4. (w, b) and (w', b') are Γ_{GELU}-separated if $(w, b) \neq (w', b')$ and $(w, b) \neq (-w', -b')$.

Definition 1.5. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called *uncoupling* if for any m and n and any family of non-degenerate and pairwise Γ_σ-separated pre-activations $\{(w_i \in \mathbb{R}^m, b_i)\}_{i=1}^n$ we have, for $a \in \mathbb{R}^n, c \in \mathbb{R}^m, d \in \mathbb{R}$,

$$\sum_{i=1}^{n} a_i \sigma(w_i \cdot x + b_i) + c \cdot x + d = 0 \quad \text{for all } x \in \mathbb{R}^m \quad \iff \quad a = 0,\ c = 0,\ d = 0. \tag{10}$$

The name *uncoupling* is chosen because σ supplies a way to go from linearly dependent pre-activations to linearly independent functions computed by rows of a shallow neural network. Our notion of uncoupling is analogous to the "independence property" of Albertini et al. (1993), the notion of "distinguishing and not affinely recursive" of Kůrková & Kainen (1994) and Definition 1(ii) in Vlačić & Bölcskei (2021), but these prior works do not include the linear term $c \cdot x$.

Proposition 1.6. *If σ is uncoupling, then so is any affine shift $\varsigma(t) = \gamma\sigma(t) + \delta t + \epsilon$, for $\gamma \neq 0$.*

Example 1.7. By Proposition 1.6, any uncoupling activation with linear odd part also has uncoupling even part. For example, since ReLU, GELU, and SiLU are uncoupling, so are the functions $x \mapsto |x|$, $x \mapsto x\,\mathrm{erf}(x/\sqrt{2})$, and $x \mapsto x\tanh(x/2)$.

Theorem 1.8. *Table 1 is correct and all activations in Table 1 are uncoupling.*

To prove that σ is uncoupling, we use a proof strategy inspired by the literature on Sigmoid, Tanh and other bounded activations (Sussmann, 1992; Albertini et al., 1993; Kůrková & Kainen, 1994), and ReLU (Agrawal & Ostrowski, 2022). We are not aware of the remaining activations in Table 1 being covered in the literature, no doubt partially because these activations were not in vogue in the 90's. We have aimed to abstract out the common properties of the proofs for all the activation functions. Somewhat oversimplified, we first show that the linear independence condition in (10) is implied by the 1D version of the same linear independence, by considering x on a carefully chosen affine line. We then apply a linear "detector" functional L on the sum in the 1D version of (10), such that L is zero on affine functions and all terms of the sum, except at a specific index i_0. Thus

we get $a_{i_0} = 0$ when the sum is zero and it follows inductively that all a_i are zero. L is chosen given an activation function and a set of pre-activations. For GELU, L "detects" the term in the sum whose second derivative decays slowest, which is unique due to the Γ_σ-separatedness of the pre-activations. The full proof is in Appendix B.

Remark 1.9. Not all activations are uncoupling. It is easy to show that $\sigma = \sin$ is not, and polynomial activations are not either. Identifiability results for polynomial activations are formulated with a dependency between the degree of the polynomial and the width of the network (Kohn et al., 2026). It could therefore be fruitful to define a notion of *uncoupling at width* N (meaning that (10) holds for $n \leq N$) to cover polynomials as well.

In order to obtain more readable formulas, it will be helpful later to pass from a general σ to a "centered" version $\tilde{\sigma}$ when possible. We make the following definition.

Definition 1.10. $\sigma : \mathbb{R} \to \mathbb{R}$ is called *centered* if its intrinsic residuals r_σ and s_σ are zero. σ is called *centerable* if there exists an affine shift $\tilde{\sigma}(t) = \gamma\sigma(t) + \delta t + \epsilon$, $\gamma \neq 0$ such that $\tilde{\sigma}$ is centered.

Proposition 1.11. *If σ is centerable to $\tilde{\sigma}(t) = \sigma(t) + p_\sigma t + q_\sigma$, then $\Gamma_\sigma = \Gamma_{\tilde{\sigma}}$, $\chi_\sigma = \chi_{\tilde{\sigma}}$ and*

$$r_\sigma(\alpha, \beta) = p_\sigma(\chi_\sigma(\alpha, \beta) - \alpha), \qquad s_\sigma(\alpha, \beta) = -p_\sigma\beta + q_\sigma(\chi_\sigma(\alpha, \beta) - 1). \tag{11}$$

Furthermore, all activations in Table 1 are centerable, with $\tilde{\sigma}(t) = \sigma(t) + p_\sigma t + q_\sigma$ centered, where p_σ and q_σ are listed in the table.

Example 1.12. $(\mathrm{GELU}(t) + \mathrm{GELU}(-t))/2 = \mathrm{GELU}(t) - t/2$, so GELU is centerable to an even activation via $p_\sigma = -1/2$ and $q_\sigma = 0$.

Remark 1.13. Not all uncoupling activations are centerable, one counterexample is $\sigma(t) = t \log |t|$.

1.2 Intrinsic Symmetries of Shallow Neural Networks

In this section we prove the existence and uniqueness of a canonical form that every shallow network (Definition 0.3) can be reduced to, which effectively removes all symmetries from the network.

Definition 1.14. An *almost canonical* shallow neural network from $\mathbb{R}^m$ to $\mathbb{R}^p$ with activation σ and width n, is a shallow neural network such that the parameters (W, b, A, C, d) satisfy

1. no column of A is zero,
2. the n pre-activations (W_i, b_i) are non-degenerate,
3. the n pre-activations are pairwise Γ_σ-separated.

Definition 1.15. Given $\Gamma \leq \mathrm{Aff}_1$, a Γ-*normalizer* is a function η taking non-degenerate pre-activations to Γ, such that $\eta(w, b) \cdot (w, b) = \eta(w', b') \cdot (w', b')$ if and only if (w, b) and (w', b') are in the same Γ-orbit. A pre-activation (w, b) is called η-*normalized* if $\eta(w, b) \cdot (w, b) = (w, b)$.

For any $\Gamma \leq \mathrm{Aff}_1$ we obtain a Γ-normalizer η by simply choosing a representative (w^*, b^*) of each Γ-orbit of pre-activations and defining $\eta(w, b)$ as an $(\alpha, \beta) \in \Gamma$ that maps (w, b) to (w^*, b^*). Possible Γ_σ-normalizers for all activations in Table 1 are presented in Table 3 in the Appendix.

Example 1.16. For GELU, a Γ_σ-normalizer is given by $\eta(w, b) = (\mathrm{sign}(w_\iota), 0)$, where w_ι is the first non-zero element of w (which exists by non-degeneracy). η-normalized pre-activations $w_\iota > 0$.

We are now ready to define the canonical form of a shallow network.

Definition 1.17. Given σ and a Γ_σ-normalizer η, a shallow neural network with activation σ is called η-*canonical* if it is almost canonical and

4. all n pre-activations are η-normalized,
5. the n pre-activations are sorted lexicographically.

Theorem 1.18. *Given a shallow neural network of width n with activation σ and a Γ_σ-normalizer η, there exists an η-canonical shallow neural network of width $n' \leq n$ computing the same function.*

The basic idea of the proof is to η-normalize all pre-activations via exploiting the symmetry of σ, delete all neurons with zero contribution to the output and finally sort the pre-activations lexicographically. Theorem 1.18 does not require σ to be an uncoupling activation. However, we will use uncoupling as a sufficient condition for uniqueness of the canonical form next.

Theorem 1.19. *Given an uncoupling activation σ and a Γ_σ-normalizer η, any two η-canonical shallow networks that define the same function must have the same parameters.*

Proving Theorem 1.19 is swift because the condition of σ being uncoupling is so strong. We subtract one network evaluation from the other and show that all coefficients must be zero, from which identification of the parameters follows. The uniqueness of the canonical form also enables us to obtain a unique *gauge* between any two almost canonical networks computing the same function. It is this gauge that will describe the weight-tying constraints of equivariant networks in Section 2.

Definition 1.20. A *network gauge* for shallow networks of width n with non-affine activation σ, consists of $\gamma \in (\Gamma_\sigma)^n$ viewed as a "column" of n values in Γ_σ and a permutation matrix π of size $n \times n$. Let $\gamma_i = (\alpha_i, \beta_i)$ and define

$$P_{(\gamma,\pi)} = \mathrm{diag}(\alpha)\pi, \qquad K_{(\gamma,\pi)} = \mathrm{diag}(\chi_\sigma(\gamma))\pi, \qquad R_{(\gamma,\pi)} = \mathrm{diag}(r_\sigma(\gamma))\pi, \tag{12}$$

so that the intrinsic symmetry of σ implies that $\sigma(P_{(\gamma,\pi)}x + \beta) = K_{(\gamma,\pi)}\sigma(x) + R_{(\gamma,\pi)}x + s_\sigma(\gamma)$. The parameter transformation $\mathcal{T}_{(\gamma,\pi)}$ defined by the gauge (γ, π) is given by

$$\begin{gathered}\mathcal{T}_{(\gamma,\pi)}(W, b, A, C, d) = (W', b', A', C', d'), \qquad \text{with} \\ W' = P_{(\gamma,\pi)}W, \qquad b' = P_{(\gamma,\pi)}b + \beta, \qquad A' = AK^{-1}_{(\gamma,\pi)}, \\ C' = C - AK^{-1}_{(\gamma,\pi)}R_{(\gamma,\pi)}W, \qquad d' = d - AK^{-1}_{(\gamma,\pi)}(R_{(\gamma,\pi)}b + s_\sigma(\gamma)).\end{gathered} \tag{13}$$

Example 1.21. If σ is centerable, then $R_{(\gamma,\pi)} = p_\sigma(K_{(\gamma,\pi)} - P_{(\gamma,\pi)})$ and $s_\sigma(\gamma) = -p_\sigma\beta + q_\sigma(K_{(\gamma,\pi)}\mathbf{1} - \mathbf{1})$. A short computation shows that the two last equations of (13) reduce to

$$C' - p_\sigma A'W' = C - p_\sigma AW, \qquad d' - p_\sigma A'b' - q_\sigma A'\mathbf{1} = d - p_\sigma Ab - q_\sigma A\mathbf{1}. \tag{14}$$

The next two propositions are crucial, but proven simply by tracing through definitions.

Proposition 1.22. *The network obtained by applying a network gauge defines the same function as the original network.*

Proposition 1.23. *The gauge transformations form a group with composition law*

$$(\gamma_2, \pi_2)(\gamma_1, \pi_1) = (\gamma_2 \odot (\pi_2\gamma_1), \pi_2\pi_1), \tag{15}$$

where $\odot$ denotes coordinate-wise product in Γ_σ. The gauge transformation in (13) *is a left action of this group on network parameters.*

The group of gauge transformations can be identified as $\Gamma_\sigma \wr S_n$, the wreath product of Γ_σ with the permutation group on n elements[1]. This group, which we will call the *gauge group*, can therefore be viewed as the intrinsic symmetry of a shallow network, stemming from the intrinsic symmetry of its activation function. For all activations in Table 1, except Exp, the gauge group is a subgroup of the group of monomial matrices (scaled permutation matrices). The only activations in Table 1 where the gauge group consists of only permutations are ELU and SELU.

Finally, we obtain gauge uniqueness for almost canonical networks from the uniqueness of the canonical form and the fact that going from an almost canonical network to its canonical form corresponds precisely to the application of a network gauge: η-normalizing gives one γ_i per pre-activation, sorting the pre-activations gives a permutation π and no neurons are deleted when transforming an almost canonical network to an η-canonical one.

Proposition 1.24. *Let σ be an uncoupling activation and assume that we are given two almost canonical shallow neural networks with σ activation that define the same function. Then the two networks have the same width and there exists a unique gauge transforming the first to the second.*

Example 1.25 (Example 0.2 continued). The layerwise neural network (4) in Example 0.2 is not almost canonical since $-W_2 = -(-a) = a = W_1$. However, by (2) we can write $f_\theta(x) = 2b\sigma(ax) - abx$, which is an almost canonical network with width $n = 1$, $A = 2b$, $W = a$, and $C = -ab$. The exotic network (5) is almost canonical with width $n = 2$, when $a \neq -b$. Thus, Proposition 1.24 says that they realize different functions, since they have different widths.

Remark 1.26. The gauge group *contains* the intertwiner group $G_{\sigma,n}$ of an uncoupling activation σ as defined by Godfrey et al. (2022). $G_{\sigma,n}$ consists of all elements $M \in \mathrm{GL}(n, \mathbb{R})$ such that there is an $N \in \mathrm{GL}(n, \mathbb{R})$ with $\sigma \circ M = N \circ \sigma$. We can show $G_{\sigma,n} \leq \Gamma_\sigma \wr S_n$ by taking any $M \in G_{\sigma,n}$

[1]Recall that $\Gamma_\sigma \wr S_n = (\Gamma_\sigma)^n \rtimes S_n$, where S_n acts on $(\Gamma_\sigma)^n$ by permuting the factors. This corresponds precisely to the composition law (15).

and looking at the two almost canonical networks $W = I, b = 0, A = I, C = 0, d = 0$ and $W' = M, b' = 0, A' = N^{-1}, C' = 0, d' = 0$, which define the same function. Thus the networks are related by a gauge (γ, π), which can be read off from $P_{(\gamma,\pi)} = \mathrm{diag}(\alpha)\pi = M, \beta = 0$. For instance, for GELU the gauge group consists of signed permutations while the intertwiner group only consists of permutations. In particular, we argue that the gauge group is a more precise notion of the symmetries induced by the activation function on the network.

2 Equivariant Shallow Neural Networks

In this section, let a group G and representations

$$\rho_{\mathrm{in}} : G \to \mathrm{GL}(m, \mathbb{R}), \qquad \rho_{\mathrm{out}} : G \to \mathrm{GL}(p, \mathbb{R}), \tag{16}$$

be given. G should not be confused with the intrinsic symmetry group Γ_σ of an activation function, G should instead be thought of as an arbitrary group that stems from a symmetry of the task at hand. Recall that equivariance of a function $f : \mathbb{R}^m \to \mathbb{R}^p$ means $f(\rho_{\mathrm{in}}(g)x) = \rho_{\mathrm{out}}(g)f(x)$.

The *adjunction trick* by Shahverdi et al. (2026a) now comes down to combining identifiability with the observation that if (W, b, A, C, d) defines an equivariant function, then the two networks $(W, b, \rho_{\mathrm{out}}(g)A, \rho_{\mathrm{out}}(g)C, \rho_{\mathrm{out}}(g)d)$ and $(W\rho_{\mathrm{in}}(g), b, A, C\rho_{\mathrm{in}}(g), d)$ define the same function for all $g \in G$. Using equivariance, almost canonicality and the uniqueness of the network gauge we obtain the following theorem and corollary.

Theorem 2.1. *Let (W, b, A, C, d) be the parameters of an almost canonical shallow network with uncoupling activation σ. Assume that the network defines a function that is equivariant with respect to ρ_{in} and ρ_{out}. Then there exists a unique homomorphism $\Psi : G \to \Gamma_\sigma \wr S_n$ such that the following weight-tying equation holds:*

$$\mathcal{T}_{\Psi(g)}(W, b, \rho_{\mathrm{out}}(g)A, \rho_{\mathrm{out}}(g)C, \rho_{\mathrm{out}}(g)d) = (W\rho_{\mathrm{in}}(g), b, A, C\rho_{\mathrm{in}}(g), d). \tag{17}$$

Definition 2.2. Two homomorphisms $\Psi : G \to \Gamma_\sigma \wr S_n$ and $\Psi' : G \to \Gamma_\sigma \wr S_n$ are said to be *conjugate* if there exists $(\gamma, \pi) \in \Gamma_\sigma \wr S_n$ such that for all $g \in G$

$$\Psi'(g) = (\gamma, \pi)\Psi(g)(\gamma, \pi)^{-1}. \tag{18}$$

Corollary 2.3. *Let two equivariant shallow almost canonical networks with uncoupling activation σ be given. Let Ψ and Ψ' be the unique homomorphisms specifying the weight-tying of the two networks, as in* (17)*. If Ψ and Ψ' are not conjugate, then the two networks define different functions.*

Example 2.4. When the intrinsic symmetry Γ_σ is trivial, as is the case for $\sigma = \mathrm{ELU}$, then (17) says that all equivariant shallow networks, after having been reduced to almost canonical form, are layerwise equivariant with a permutation action on the intermediate layer:

$$\begin{gathered} W\rho_{\mathrm{in}}(g) = \pi(g)W, \qquad b = \pi(g)b, \qquad A\pi(g) = \rho_{\mathrm{out}}(g)A, \\ C\rho_{\mathrm{in}}(g) = \rho_{\mathrm{out}}(g)C, \qquad d = \rho_{\mathrm{out}}(g)d. \end{gathered} \tag{19}$$

Example 2.5. For centerable σ, (14) and (17) yield that any shallow equivariant network reduced to almost canonical form satisfies

$$\begin{gathered} W\rho_{\mathrm{in}}(g) = P_{\Psi(g)}W, \qquad b = P_{\Psi(g)}b + \beta(g), \qquad AK_{\Psi(g)} = \rho_{\mathrm{out}}(g)A, \\ (C - p_\sigma AW)\rho_{\mathrm{in}}(g) = \rho_{\mathrm{out}}(g)(C - p_\sigma AW), \\ d - p_\sigma Ab - q_\sigma A\mathbf{1} = \rho_{\mathrm{out}}(g)(d - p_\sigma Ab - q_\sigma A\mathbf{1}). \end{gathered} \tag{20}$$

To give some intuition here, if σ is centered ($p_\sigma = q_\sigma = 0$), these become layerwise linear equivariance constraints, but with the representation on the latent space changing from $P_{\Psi(g)}$ to $K_{\Psi(g)}$ when passing through σ. This shows that passing from a non-centered σ to its centered version $\tilde{\sigma}$ gives an easy way to parameterize exotic σ-networks. In fact, the construction by Agrawal & Ostrowski (2023) to parameterize deep ReLU-networks with signed permutation representations on the hidden layers implicitly rewrites the networks in terms of $\tilde{\sigma}(t) = |t|$.

Example 2.6. Consider the case $\sigma = \mathrm{GELU}$. Then $\beta = 0$, $\chi_\sigma = 1$ and σ is centerable with $p_\sigma = -1/2$ and $q_\sigma = 0$, yielding the weight-tying equations

$$\begin{gathered} W\rho_{\mathrm{in}}(g) = \mathrm{diag}(\alpha(g))\pi(g)W, \qquad b = \mathrm{diag}(\alpha(g))\pi(g)b, \qquad A\pi(g) = \rho_{\mathrm{out}}(g)A, \\ (C + AW/2)\rho_{\mathrm{in}}(g) = \rho_{\mathrm{out}}(g)(C + AW/2), \qquad (d + Ab/2) = \rho_{\mathrm{out}}(g)(d + Ab/2), \end{gathered} \tag{21}$$

where $\mathrm{diag}(\alpha(g))\pi(g)$ is a signed permutation representation.

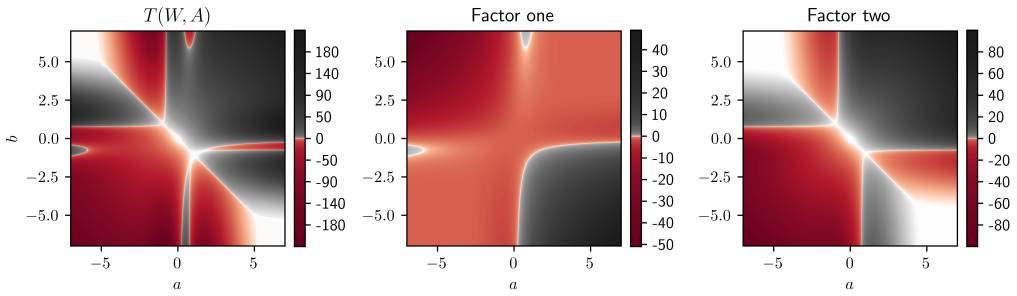

Figure 3: The contour plot of $T(W, A)$ (left), the factor $\big(\Phi(b) - \Phi(a)\big) - 1$ (middle), and the factor $ab\big(b\varphi(a) + a\varphi(b)\big) + (a^2 + b^2)\big(\Phi(b) - \Phi(-a)\big)$ (right). The zeros of factor one come from zero loss solutions. Note that $a = -b$ gives zeroes because in that configuration the network is layerwise equivariant. Finally, special configurations cause the sum in factor two to be 0.

Corollary 2.3 makes it relevant to classify all homomorphisms $\Psi : G \to \Gamma_\sigma \wr S_n$ up to conjugation, because this gives all ways that networks can be equivariant: each network "type" is obtained by picking a Ψ and solving (17) for almost canonical solutions. The classification of homomorphisms into a wreath product is known in the literature (Kovács, 1985; 1988), we state the relevant result in our notation in Proposition B.11. For given Ψ, ρ_{in} and ρ_{out} it is possible that no almost canonical equivariant shallow network from ρ_{in} to ρ_{out} exists with Ψ as its homomorphism and a general way to determine this seems difficult to find, beyond solving (17). For invariant ReLU-networks with G finite and ρ_{in} orthogonal, (Agrawal & Ostrowski, 2022, Theorem 4(a)) gives a group-theoretic condition for realizability, that could potentially serve as a starting point for generalization.

We provide a worked example of how to use the theory in this section to design neural networks in Appendix A, where we cover two layerwise and two exotic parameterizations of MLP blocks in equivariant vision transformers. We will use those constructions in the experiments in Section 4.

Remark 2.7. In practice, we care most about the cases $C = 0$ or $C = I$, i.e. no linear skip connection or identity skip connection. For $C = 0$, (20) gives $p_\sigma AW\rho_{\text{in}}(g) = p_\sigma \rho_{\text{out}}(g)AW$, which is vacuous for $p_\sigma = 0$ and an equivariance constraint for AW otherwise. For $C = I$ and $\rho_{\text{in}} = \rho_{\text{out}}$, we get the same constraint. Crucially, $C = 0$ is not preserved when passing from an activation to its centered version, so analogues of the ReLU-networks by Agrawal & Ostrowski (2023) are not available in this case. Furthermore, $C = 0$ is not preserved under gauge transformations in general. Thus, the classification of Ψ up to conjugation does not yield a classification of equivariant networks that additionally satisfy $C = 0$. We can however still use our theory to design equivariant networks via the obtained constraints, we will see a practical example in Appendix A.

3 Optimization Properties

For orthogonal representations, the space of layerwise equivariant parameters is known to be invariant under gradient flow with full group orbit data augmentation (Nordenfors et al., 2025). Example 0.2 provides a counterexample for exotic equivariant parameters. Let $f_\theta = A\sigma(Wx)$ be a shallow neural network with activation $\sigma = \text{GELU}$, parametrized by $\theta = (W \in \mathbb{R}^{2\times 1}, A \in \mathbb{R}^{1\times 2})$. As we have seen, f_θ is even if and only if $AW = 0$. Suppose that we have i.i.d. data (x, y), where x is Rademacher, $P(x = 1) = P(x = -1) = 1/2$, ground truth $y = x^2$, and MSE loss $\ell(f, y) = (f - y)^2$. The augmented risk is $R^{\text{aug}}(\theta) = \mathbb{E}_{x,y}[\frac{1}{2}(\ell(f_\theta(x), y) + \ell(f_\theta(-x), y))]$ and the flow under consideration is

$$\dot{W} = -\nabla_W R^{\text{aug}}(\theta(t)), \qquad \dot{A} = -\nabla_A R^{\text{aug}}(\theta(t)). \tag{22}$$

We thus want to know whether $T(W, A) := A\dot{W} + \dot{A}W = 0$, when $AW = 0$, since this is the condition for equivariance to be maintained under the flow. A short calculation (see Appendix D.2) shows that if A and W satisfy the exotic equivariance constraint (5), then

$$T(W, A) = -2\Big(ab\big(\Phi(b) - \Phi(a)\big) - 1\Big)\Big(ab\big(b\varphi(a) + a\varphi(b)\big) + (a^2 + b^2)\big(\Phi(b) - \Phi(-a)\big)\Big), \tag{23}$$

where Φ and φ are the standard normal CDF and PDF, respectively. As can be seen from the contour plot in Figure 3, $T(W, A)$ is far from being everywhere 0. For example, $T(W, A) > 0$ for $0 < b < a$.

4 VISION TRANSFORMER EXPERIMENTS

We conduct a proof-of-concept experiment to highlight the impact of the choice of Ψ when designing equivariant networks. We modify the MLP blocks in the horizontal reflection equivariant vision transformers by Bökman et al. (2025) to satisfy weight-tying according to four different Ψ_i that are described in detail in Appendix A. In these experiments, $G \cong \mathbb{Z}_2$ and $\sigma = \mathrm{GELU}$.

Ψ_1, Ψ_2 are two different layerwise constructions, with Ψ_1 containing copies of the regular representation only, while Ψ_2 contains half channels with the regular representation and the remaining channels having the trivial representation. Ψ_3, Ψ_4 are two different exotic equivariant constructions. Our theory shows that these four choices lead to MLP blocks that can realize different functions. The "residual stream" of the transformer carries the same representation in all cases, namely a multiple of the regular representation, so $\rho_{\mathrm{in}} = \rho_{\mathrm{out}}$ is the same for all.

Table 2: ViT-L on ImageNet-1K

Equivariance constraint	Acc@1	Acc@5
—	76.0%	92.3%
Ψ_1	74.3%	91.5%
Ψ_2	**76.2%**	**92.4%**
Ψ_3	73.2%	90.7%
Ψ_4	71.8%	90.2%

We train size Large vision transformers on ImageNet-1K, using images of resolution 96×96 (to reduce compute) and the DeiT-III training recipe. This recipe includes horizontal reflection augmentations, so the task is equivariant. Details on the experimental setup are reported in Appendix A.2. Table 2 shows the obtained validation accuracies. We report the maximum score over five seeds, all seeds are reported in the appendix. We wish to make four main points here. First, the choice of Ψ matters. Second, the exotic Ψ_3 and Ψ_4 are not the best in this experiment, which is good news for practitioners since they are most bothersome to implement. Third, we reproduce the result by Bökman et al. (2025) that Ψ_1 underperforms the unconstrained (non-equivariant) baseline "—". Fourth, since there are only finitely many choices of Ψ (but a lot more than four), there must be an optimal choice per MLP block, and it is probably not one of the four we tried here. Furthermore, the unconstrained network does not find the optimal Ψ per MLP block either as it performs worse than the best run of a network with $\Psi = \Psi_2$ in every block (some of the other seeds for Ψ_2 were worse).

5 CONCLUSION, LIMITATIONS, & OUTLOOK

With Theorem 2.1 we found a complete characterization of all equivariant shallow neural networks with uncoupling activations. The symmetries of a neural network form a group of gauge transformations $\Gamma_\sigma \wr S_n$ that preserve the function computed by the network. For two almost canonical networks computing the same function, we showed that the gauge transforming one to the other is unique. For every almost canonical G-equivariant network there thus exists a unique homomorphism $\Psi : G \to \Gamma_\sigma \wr S_n$ that determines the weight-tying constraints. Some of these are exotic. Beyond computing distinct equivariant functions, we showed that the different parametrizations can have different optimization dynamics. In particular, we showed that exotic parametrizations are not guaranteed to be invariant under group orbit augmented gradient flow, which is known to be the case for layerwise equivariant parametrizations. We also conducted an experiment to verify that the choice of equivariance constraint matters when designing neural networks. Within the scope of the experiment, we saw that one choice of layerwise structure Ψ_2 performed better across all seeds and its best run was even better than the best run of an unconstrained baseline (which covers all Ψ).

There are several limitations that would be interesting directions for future study. We restricted our attention to shallow networks which leaves open the characterization of equivariant deep neural networks. Secondly, we show Ψ is unique given an equivariant almost canonical network with fixed ρ_{in}, ρ_{out}, but it remains an open problem to determine when a given triple $(\Psi, \rho_{\mathrm{in}}, \rho_{\mathrm{out}})$ is realizable as an equivariant almost canonical network. For optimization, sufficient conditions are known for invariance under data augmented gradient flow, but not necessary ones. From the practitioners' point of view, the most pressing concern would be to determine which choice of Ψ is best in a specific instance. In practice, MLPs are usually given with skip $C = 0$ or $C = I$, conditions that are not preserved under gauge transformations. Thus, it is relevant to further study how to classify equivariant networks under this additional constraint. Finally, an important question is to determine under what conditions approximately equivariant networks are guaranteed to approximate the weight-tying constraints given by Theorem 2.1.

AI USE STATEMENT

We have used Codex and Claude Code (both via subscription usage) for the following research subtasks, with required disclosure: *help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, implement methods, design or provide feedback on research methodology or experiments*, and the following tasks with recommended disclosure: *create or modify scientific figures or images, create or edit software code, creation of artifacts, summarize or analyse existing literature, brainstorming, sourcing/searching for information, identify relevant literature*.

More specifically, we have used the LLMs for

- Theorem proving and Lean formalization of results. Initially, author XY used LLMs to formalize proofs for the identifiability of GELU, SiLU and Softplus shallow networks, as well as the generalization of the Agrawal & Ostrowski (2022) canonical form and consequences for equivariant networks, that we had obtained by hand. With this formalization as a starting point, XY with LLM assistance went through many rounds of iteration to get the proofs as streamlined and to XY's liking as possible. It became clear that it was possible to fit many modern activations into a single proof route similar to the one used by Albertini et al. (1993), as presented in Appendix B.1. Once this proof route was formalized, it was possible to prompt LLMs to add further activations. For instance, the "detector functionals" used for ReLU, Exp, and ELU were provided by an LLM with access to the Lean code. XY found it a slightly dizzying experience to have a Lean code base to essentially talk and ask questions to via an LLM, but it was also quite useful and interesting.
- Understanding the wreath product theory of Kovács and formalizing it in Lean. The proof in Appendix B.4 is human written.
- Python pair programming for the ViT experiments.
- Generating the tikz-plots in this manuscript.
- Reviewing the manuscript, recommending additional related work, checking for typos and left out arguments etc.

We have reviewed all AI-assisted work: LLM generated proofs were checked and rewritten by hand, Lean statements were reviewed for alignment with the intended meaning and checked via Comparator. Python code was checked by eye and via LLM generated tests.

We have not used generative AI tools for *writing this manuscript*, nor any of these tasks with required or recommended disclosure: *propose or refine hypotheses, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, suggest experimental parameters, draft parts of a research paper, interpret results, discover research topics or identify gaps, edit a research paper to improve readability, format references, suggest a structure for a research paper, propose a title or keywords for a research paper.*

These tasks with required or recommended disclosure are not applicable to this work: *generate synthetic data sets, formulate questions for surveys or interviews, transcribe recordings of research material.*

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release our code open source upon acceptance of the paper to the conference.

REFERENCES

Devanshu Agrawal and James Ostrowski. A classification of g-invariant shallow neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13679–13690. Curran Associates, Inc., 2022. doi: 10.52202/068431-0994. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/58b9a640af6d69781e90969d936e87ce-Paper-Conference.pdf.

Devanshu Agrawal and James Ostrowski. Densely connected g-invariant deep neural networks with signed permutation representations. *Journal of Machine Learning Research*, 24(370):1–40, 2023. URL http://jmlr.org/papers/v24/23-0294.html.

Francesca Albertini, Eduardo D. Sontag, and Vincent Maillot. Uniqueness of weights for neural networks. In Richard J. Mammone (ed.), *Artificial Neural Networks for Speech and Vision*, volume 4 of *Chapman & Hall Neural Computing Series*, pp. 113–125, London, 1993. Chapman and Hall. URL https://sontaglab.org/FTPDIR/92caip.pdf.

Georg Bökman and Fredrik Kahl. Investigating how relu-networks encode symmetries. *Advances in Neural Information Processing Systems*, 36:13720–13744, 2023.

Georg Bökman, David Nordström, and Fredrik Kahl. Flopping for FLOPs: Leveraging equivariance for computational efficiency. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4823–4838. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bokman25a.html.

Joachim Bona-Pellissier, François Bachoc, and François Malgouyres. Parameter identifiability of a deep feedforward relu neural network. *Machine learning*, 112(11):4431–4493, 2023.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017. doi: 10.1109/MSP.2017.2693418.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3):507–555, 1994. URL http://eudml.org/doc/39464.

Bella Finkel, Jose Israel Rodriguez, Chenxi Wu, and Thomas Yahl. Activation degree thresholds and expressiveness of polynomial neural networks. *Algebraic Statistics*, 16(2):113–130, Sept 2025. ISSN 2693-2997. doi: 10.2140/astat.2025.16.113. URL http://dx.doi.org/10.2140/astat.2025.16.113.

Johanna Marie Gegenfurtner, Moritz Grillo, and Guido Montúfar. The symmetries of three-layer relu networks, 2026. URL https://arxiv.org/abs/2605.18319.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. *Advances in Neural Information Processing Systems*, 35:11893–11905, 2022.

J. Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden symmetries of relu networks. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Semlaflow – efficient 3d molecular generation with latent attention and equivariant flow matching. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=bee2G6pEh0.

Mohammad Mohaiminul Islam, Rishabh Anand, David R Wessels, Friso de Kruiff, Thijs P. Kuipers, Rex Ying, Clara I. Sánchez, Sharvaree Vadgama, Georg Bökman, and Erik J Bekkers. Platonic transformers: A solid choice for equivariance. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JTu5Mawn86.

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a0dc078ca0d99b5ebb465a9f1cad54ba-Paper.pdf.

Kathlén Kohn, Giovanni Luca Marchetti, Alex Massarenti, and Massimiliano Mella. Linear independence of polynomial compositions and identifiability of deep neural networks, 2026. URL https://arxiv.org/abs/2608.27113.

LG Kovács. Two results on wreath products. *Archiv der Mathematik*, 45(2):111–115, 1985.

LG Kovács. Some theorems on wreath-products. *Publicationes Mathematicae-Debrecen*, 35(1-2): 155–160, 1988.

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

Giovanni Luca Marchetti, Christopher J Hillar, Danica Kragic, and Sophia Sanborn. Harmonics of learning: Universal fourier features emerge in invariant networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 3775–3797. PMLR, 2024.

Flavio Martinelli, Berfin Simsek, Wulfram Gerstner, and Johanni Brea. Expandand-cluster: Parameter recovery of neural networks. In *International Conference on Machine Learning*, pp. 34895–34919, 2023.

Oskar Nordenfors, Fredrik Ohlsson, and Axel Flinth. Optimization dynamics of equivariant and augmented neural networks. *Transactions on Machine Learning Research*, 2025.

Vahid Shahverdi, Giovanni Luca Marchetti, Georg Bökman, and Kathlén Kohn. Identifiable equivariant networks are layerwise equivariant. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=32Us54GZlT.

Vahid Shahverdi, Giovanni Luca Marchetti, and Kathlén Kohn. Learning on a razor's edge: Identifiability and singularity of polynomial neural networks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 26948–26970, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2d1351d0258e4469e7bb0c1812372960-Paper-Conference.pdf.

Héctor J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(05)80037-1. URL https://www.sciencedirect.com/science/article/pii/S0893608005800371.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pp. 516–533. Springer, 2022.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Verner Vlačić and Helmut Bölcskei. Affine symmetries and neural network identifiability. *Advances in Mathematics*, 376:107485, 2021. ISSN 0001-8708. doi: https://doi.org/10.1016/j.aim.2020.107485. URL https://www.sciencedirect.com/science/article/pii/S0001870820305132.

Jeffrey Wood and John Shawe-Taylor. Representation theory and invariant neural networks. *Discrete Applied Mathematics*, 69(1):33–60, 1996. ISSN 0166-218X. doi: https://doi.org/10.1016/0166-218X(95)00075-3. URL https://www.sciencedirect.com/science/article/pii/0166218X95000753.

Yani Zhang and Helmut Bölcskei. Complete identification of deep relu networks through łukasiewicz logic, 2026. URL https://arxiv.org/abs/2602.00266.

Table 3: Γ_σ-normalizers η for common activations σ. η is not unique given Γ_σ, we list possible ones. We write (w^*, b^*) for the normalized pre-activation $\eta(w,b)\cdot(w,b)$. w_ι is the first non-zero element of w, which must exist since (w,b) is assumed to be non-degenerate.

σ	Γ_σ	$\eta(w,b)$	η-normalized (w^*,b^*)
ReLU	$\beta=0$	$(1/w_\iota, 0)$	$w^*_\iota = 1$
Tanh, Sigmoid, GELU, SiLU, Softplus	$\alpha=\pm1, \beta=0$	$(\mathrm{sign}(w_\iota), 0)$	$w^*_\iota > 0$
Exp	$\alpha=1$	$(1,-b)$	$b^*=0$
ELU, SELU	$\alpha=1,\beta=0$	$(1,0)$	Any (w,b) is normalized

A Worked Practical Example: Vision Transformer MLPs

In recent equivariant vision transformers, the feature spaces are multiple copies of the regular representation (Bökman et al., 2025; Islam et al., 2026). In transformers, the MLP blocks are shallow networks, making the theory developed in this paper apply directly.[2] Let's consider the case of the (multiplicative) group $G = \{1,-1\} \cong \mathbb{Z}_2$, representing image reflections, and $\rho_{\text{in}}(-1) = \rho_{\text{out}}(-1) = (1)^{\oplus m/2} \oplus (-1)^{\oplus m/2}$, $m/2$ copies of the regular representation of G in the irrep basis. We consider $\sigma = \text{GELU}$ as used by Bökman et al. (2025), and many common transformer implementations. We obtain $\Gamma_\sigma \cong G = \{1,-1\}$.

Let us determine all possible choices for weight-tying constraints, i.e. choices of homomorphisms $\Psi : G \to \Gamma_\sigma \wr S_n$. This is the content of Proposition B.11. Proposition B.11 says that we should first find all homomorphisms from subgroups of G to Γ_σ. G has only two subgroups, itself and the trivial group. There are only three possible homomorphisms:

$$\phi_0 : \{1\} \to \Gamma_\sigma, \quad \phi_1 : G \to \Gamma_\sigma, \phi_1(-1) = 1, \text{ and } \phi_{-1} : G \to \Gamma_\sigma, \phi_{-1}(-1) = -1. \tag{24}$$

They are pairwise non-equivalent. So, according to Proposition B.11, the choice of Ψ consists of the following:

1. Select how many of the $4m$ hidden dimension should form permutation orbits of size 1 and 2 respectively, which determines π.
2. For each size 1 orbit, choose one of ϕ_1 and ϕ_{-1}. This determines γ. There is no choice to be made for the size 2 orbits, since only ϕ_0 exists.

All of these choices of Ψ are realizable by almost canonical networks by the following construction. By taking just one input pair (x_1, x_{-1}) transforming according to $(1)\oplus(-1)$, we can map to size 1 orbits of type ϕ_1 via $(x_1,x_{-1}) \mapsto x_1$, to size 1 orbits of type ϕ_{-1} via $(x_1,x_{-1}) \mapsto x_{-1}$ and to size 2 orbits via $(x_1,x_{-1}) \mapsto (x_1 + kx_{-1}, x_1 - kx_{-1})$ with $k>1$, say. These are all Γ_σ-separated pre-activations (with $b=0$). Arbitrarily many such orbits/pre-activations can be made Γ_σ-separated by simply scaling them differently. To select A, C, d such that (21) holds, we can do as follows. A can be chosen as having first row consisting of only 1's, which is equivariant under any permutation representation, mapping to an invariant output. The remaining rows of A can be set to 0, C can be set to $-AW/2$ and d to 0.

In practice, the shallow networks inside the MLP blocks in transformers do not include a linear skip connection C. Instead they have trivial linear skip connection, i.e. $C=0$. What do such equivariant MLPs look like? They need to satisfy $C=0$ and (21).

Plugging $C=0$ into (21), we obtain that the extra condition is

$$AW\rho_{\text{in}}(g) = \rho_{\text{out}}(g)AW. \tag{25}$$

We note that since every Ψ is realizable in an almost canonical net for $\sigma = \text{GELU}$ as explained earlier, every $\Psi = \Psi'^{\oplus 2}$, consisting of two copies of the same Ψ', is realizable in an almost canonical

[2] The MLP blocks are surrounded by identity skip connections, and include also LayerNorm and potentially LayerScale layers, but those are equivariant under $\rho_{\text{in}} = \rho_{\text{out}}$ by default in equivariant transformers and not relevant here. We focus on the core shallow network of the MLP block.

net that additionally satisfies $C = 0$. To show this, note that if A', W', b' realize Ψ', then

$$W = \begin{pmatrix} W' \\ tW' \end{pmatrix}, \quad b = \begin{pmatrix} b' \\ tb' \end{pmatrix}, \quad A = \begin{pmatrix} tA' & -A' \end{pmatrix}, \tag{26}$$

realize Ψ, and satisfy $AW = 0$, $Ab = 0$ so that (21) holds with $C = 0, d = 0$. By selecting t we can make all pre-activations Γ_{GELU}-separated. Our choices of Ψ below will be of this "duplicated" form, so all of them admit almost canonical network realizations with $C = 0$. We will however not parameterize them like (26) because it is not the most general form for networks (almost canonical or otherwise) realizing Ψ.

In the experiments we will consider the following four choices of Ψ, with $S = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (we now write $\Psi(g) = P_{\Psi(g)} = \mathrm{diag}(\alpha(g))\pi(g)$ to simplify notation).

$$\begin{aligned} \Psi_1(-1) &= S^{\oplus 2m}, \qquad \Psi_2(-1) = (1)^{\oplus 2m} \oplus S^{\oplus m}, \\ \Psi_3(-1) &= (1)^{\oplus 2m} \oplus (-1)^{\oplus 2m}, \qquad \Psi_4(-1) = -S^{\oplus 2m}, \end{aligned} \tag{27}$$

where Ψ_1 and Ψ_2 are permutation representations and hence correspond to ordinary layerwise equivariant networks, whereas Ψ_3 and Ψ_4 are exotic. These all give width $4m$ networks, as is standard in transformer MLP blocks. Ψ_1 could be said to be the canonical choice, as used in the equivariant ViTs by Bökman et al. (2025). The homomorphisms Ψ_1, Ψ_2 and Ψ_3 are all pairwise non-conjugate. Ψ_4 is conjugate to Ψ_1 and hence any almost canonical network of either type can be rewritten in terms of the other. However, with the additional constraint $C = 0$, Ψ_4 and Ψ_1 are not guaranteed to be equivalent, because to go from an almost canonical network of type Ψ_4 to one of type Ψ_1 in general requires setting $C \neq 0$.

The constraints on W, b, A from (21) are ordinary linear equivariance constraints that can be solved in various ways, they are standard in geometric deep learning (Wood & Shawe-Taylor, 1996; Bronstein et al., 2021). We follow Bökman et al. (2025) for our ViT implementations and parameterize everything in the irrep domain, where Schur's lemma implies that the linear maps block-diagonalize into one block per irrep.

However, we now also need to consider the final two "exotic" constraints in (21). For Ψ_1 and Ψ_2 they are trivially upheld via equivariance of W, A, b, by just enforcing additionally the standard linear constraint $d = \rho_{\mathrm{out}}(g)d$.

For Ψ_3, we can split W into a block-diagonal matrix by Schur's lemma, i.e.

$$W = \begin{pmatrix} W_+ & 0 \\ 0 & W_- \end{pmatrix} \tag{28}$$

where W_+ and W_- are the intertwiners on irreps of type (1) and (-1) respectively. We have $\pi_3(-1) = I$, which is part of A's equivariance constraint: $\rho_{\mathrm{out}}(-1)A = A\pi_3(-1)$. Schur's lemma now implies that A only maps from irreps of type (1) to irreps of type (1), and so A is a block matrix

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & 0 \end{pmatrix}. \tag{29}$$

The condition (25) (which is an equivariance condition!) means, again by Schur, that AW itself is block-diagonal in the same form as W. But this means that $A_2W_- = 0$, which is hence an exotic constraint for Ψ_3. The constraint on d from the last equation of (21) in this case is again the standard linear constraint $\rho_{\mathrm{out}}(g)d = d$, because $\rho_{\mathrm{out}}(g)A = A$.

For Ψ_4 note that $P_4(-1) = -\pi_4(-1)$. Thus, $AW\rho_{\mathrm{in}}(-1) = -A\pi_4(-1)W = -\rho_{\mathrm{out}}(-1)AW$ by equivariance of W and A. But in combination with (25), this yields $AW = 0$ which is an exotic constraint for Ψ_4. In this case the last equation of (21) gives the further exotic constraint that the last half of d must equal the last half of $-Ab/2$.

A.1 Some Intuition About the Different Ψ_i

As mentioned, Ψ_1 can be considered the canonical choice, and it has equally many of both irrep types. However, for a classification task such as ImageNet-1K, we can speculate that invariant features (irrep type (1)), are more important. Ψ_2 contains three times as many irrep (1) features than irrep (-1) features, while still yielding a standard layerwise equivariant network.

Table 4: ViT-L on ImageNet-1K. We report the validation accuracy at the end of training over five seeds. Table 2 shows the best run per network type from this table. Those runs are bold here.

Equivariance constraint	Seed 0		Seed 1		Seed 2		Seed 3		Seed 4	
	%@1	%@5	%@1	%@5	%@1	%@5	%@1	%@5	%@1	%@5
—	75.7	92.0	**76.0**	**92.3**	75.9	92.2	76.0	92.2	75.9	92.2
Ψ_1	**74.3**	**91.5**	72.5	90.8	72.0	90.1	65.7	86.8	72.5	90.8
Ψ_2	76.1	92.5	75.9	92.3	73.6	91.1	73.7	91.3	**76.2**	**92.4**
Ψ_3	**73.2**	**90.7**	72.8	90.6	73.1	90.7	73.1	90.6	73.1	90.7
Ψ_4	70.6	89.4	**71.8**	**90.2**	67.3	87.5	68.5	88.2	70.0	89.2

Ψ_4 is a form of "evil twin" of Ψ_1, which also consists equally many of each irrep. Finally, Ψ_3 has a constraint on A, (29), which leads to only outputting type (1) features (which we speculated were most important for the classification task). Note that the MLP block skip-connections mean that type (-1) features are still preserved, but they are not updated by the MLP block shallow networks. The (-1) features are still updated in the attention blocks for Ψ_3-networks.

A.2 Experiment Details

We use the implementation by Bökman et al. (2025) as a base, which follows the DeiT-III training recipe for vision transformers (Touvron et al., 2022). To implement the exotic weight constraints, we project to the conditions $AW = 0$ every 25 optimizer steps. We hold W fixed and project A to the nullspace of W^T. If $\|AW\|/(\|A\|\|W\|) \geq \epsilon_{\text{tol}}$, where $\|\cdot\|$ is the Frobenius norm, we subsequently fix A and project W to the nullspace of A. The projection is done in FP32 precision, while the rest of the training uses autocast to BF16. We set the tolerance to 10^{-6} and never observe a failure to obtain $\|AW\|/(\|A\|\|W\|) < \epsilon_{\text{tol}}$ after both described projections. The interval of 25 steps was chosen to reduce the cost of our non-optimal projection code. There are 625 steps per epoch, so the projections are frequent and do not allow the network to stray from the constraints. We do not rule out the possibility that a more optimized implementation for optimization subject to $AW = 0$ (e.g. a re-parameterization) would yield better results.

We train for 200 epochs on ImageNet-1K (Deng et al., 2009), with image resolution 96×96, which is a choice made to reduce the training costs relative to the standard DeiT-III recipes (400 or 800 epochs at resolution 192×192 or 224×224). In our setup, training a ViT-L on two H100 GPUs takes around 22 hours.

The obtained results across five seeds are reported in Table 4. We note that the training of the unconstrained baseline seems to be more stable across different seeds. This may not be due to any concrete benefit of unconstrained training, but may as well simply be due to the fact that all the hyperparameters in the DeiT-III training recipe by Touvron et al. (2022) are tuned for the unconstrained setting. In particular, the unconstrained network has twice the number of parameters compared to our equivariant networks. This potential issue was also mentioned by Bökman et al. (2025). Retuning the hyperparameters would require significant compute, and we deem it out of scope for this paper. The two main points to be taken from Table 4 are that the different Ψ yield different performance and that the unconstrained baseline does not find an optimal equivariant structure since it is outperformed by some equivariant networks of type Ψ_2.

B Proofs and Additional Results

This section contains proofs of the results in the main paper. As supplementary material, we provide Lean code formalizing all theorems.

B.1 Activation Functions

Proposition 1.2. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-affine function. Then γ, δ, ϵ in* (8) *are unique and $\gamma \neq 0$. We can therefore define maps $\chi_\sigma : \Gamma_\sigma \to \mathbb{R} \setminus \{0\}$ and $r_\sigma : \Gamma_\sigma \to \mathbb{R}$, $s_\sigma : \Gamma_\sigma \to \mathbb{R}$ as $\chi_\sigma(\alpha, \beta) = \gamma$ and $r_\sigma(\alpha, \beta) = \delta, s_\sigma(\alpha, \beta) = \epsilon$ via* (8)*. χ_σ is called the* intrinsic character *of σ, as it is a character of Γ_σ (i.e. χ_σ is a group homomorphism). r_σ, s_σ are called the* intrinsic linear and constant residual *of σ respectively, they are not group homomorphsims in general.*

Proof. $\gamma = 0$ is excluded by the fact that σ is non-affine.

Next, assume that $\gamma_1, \delta_1, \epsilon_1$ and $\gamma_2, \delta_2, \epsilon_2$ both satisfy

$$\sigma(\alpha t + \beta) = \gamma_i \sigma(t) + \delta_i t + \epsilon_i. \tag{30}$$

If $\gamma_1 \neq \gamma_2$, then $\sigma(t) = ((\delta_2 - \delta_1)t + \epsilon_2 - \epsilon_1)/(\gamma_1 - \gamma_2)$ which contradicts the fact that σ is non-affine. So $\gamma_1 = \gamma_2$, which implies $\delta_1 t + \epsilon_1 = \delta_2 t + \epsilon_2$, meaning $\delta_1 = \delta_2$, $\epsilon_1 = \epsilon_2$.

Next we show $\chi((\alpha', \beta')(\alpha, \beta)) = \chi(\alpha', \beta')\chi(\alpha, \beta)$ Let $(\alpha'', \beta'') = (\alpha', \beta')(\alpha, \beta) = (\alpha'\alpha, \alpha'\beta + \beta')$. We have

$$\sigma(\alpha'' t + \beta'') = \chi(\alpha'', \beta'')\sigma(t) + r(\alpha'', \beta'')t + s(\alpha'', \beta'') \tag{31}$$

but also

$$\begin{aligned} \sigma(\alpha'' t + \beta'') &= \chi(\alpha', \beta')\sigma(\alpha t + \beta) + r(\alpha', \beta')(\alpha t + \beta) + s(\alpha', \beta') \\ &= \chi(\alpha', \beta')(\chi(\alpha, \beta)\sigma(t) + r(\alpha, \beta)t + s(\alpha, \beta)) + r(\alpha', \beta')(\alpha t + \beta) + s(\alpha', \beta') \end{aligned} \tag{32}$$

and uniqueness now gives the statement. □

Lemma B.1. *If σ is an activation and ς is an affine shift of σ, $\varsigma(t) = \gamma\sigma(t) + \delta t + \epsilon$, with $\gamma \neq 0$, then $\Gamma_\varsigma = \Gamma_\sigma$.*

Proof. Let σ is an activation and let $\varsigma(t) = \gamma\sigma(t) + \delta t + \epsilon$, with $\gamma \neq 0$. For $(\alpha, \beta) \in \Gamma_\varsigma$ We have

$$\begin{aligned} \varsigma(\alpha t + \beta) &= \gamma'\varsigma(t) + \delta' t + \epsilon' \\ &= \gamma'(\gamma\sigma(t) + \delta t + \epsilon) + \delta' t + \epsilon' \\ &= \gamma'\gamma\sigma(t) + (\gamma'\delta + \delta')t + (\gamma'\epsilon + \epsilon'), \end{aligned} \tag{33}$$

and

$$\varsigma(\alpha t + \beta) = \gamma\sigma(\alpha t + \beta) + \delta(\alpha t + \beta) + \epsilon \tag{34}$$

By (33) and (34) we have

$$\sigma(\alpha t + \beta) = \frac{\gamma'\gamma}{\gamma}\sigma(t) + \frac{\gamma'\delta + \delta' - \delta\alpha}{\gamma}t + \frac{\gamma'\epsilon + \epsilon' - \delta\beta - \epsilon}{\gamma} \tag{35}$$

Thus $\Gamma_\varsigma \subseteq \Gamma_\sigma$. Thus, we have shown that if ς is an affine shift of σ, then $\Gamma_\varsigma \subseteq \Gamma_\sigma$. However, if ς is an affine shift of σ, then σ is an affine shift of ς, and so $\Gamma_\sigma \subseteq \Gamma_\varsigma$. Thus, $\Gamma_\varsigma = \Gamma_\sigma$. □

Proposition 1.11. *If σ is centerable to $\tilde{\sigma}(t) = \sigma(t) + p_\sigma t + q_\sigma$, then $\Gamma_\sigma = \Gamma_{\tilde{\sigma}}$, $\chi_\sigma = \chi_{\tilde{\sigma}}$ and*

$$r_\sigma(\alpha, \beta) = p_\sigma(\chi_\sigma(\alpha, \beta) - \alpha), \qquad s_\sigma(\alpha, \beta) = -p_\sigma\beta + q_\sigma(\chi_\sigma(\alpha, \beta) - 1). \tag{11}$$

Furthermore, all activations in Table 1 are centerable, with $\tilde{\sigma}(t) = \sigma(t) + p_\sigma t + q_\sigma$ centered, where p_σ and q_σ are listed in the table.

Proof. Assume that σ is centerable to $\tilde{\sigma}(t) = \sigma(t) + p_\sigma t + q_\sigma$. By Lemma B.1 we have $\Gamma_\sigma = \Gamma_{\tilde{\sigma}}$. For $(\alpha, \beta) \in \Gamma_\sigma = \Gamma_{\tilde{\sigma}}$, we have

$$\begin{aligned} \tilde{\sigma}(\alpha t + \beta) &= \sigma(\alpha t + \beta) + p_\sigma(\alpha t + \beta) + q_\sigma \\ &= \chi_\sigma(\alpha, \beta)\sigma(t) + (r_\sigma(\alpha, \beta) + p_\sigma\alpha)t + (p_\sigma\beta + q_\sigma + s_\sigma(\alpha, \beta)) \\ &= \chi_\sigma(\alpha, \beta)(\tilde{\sigma}(t) - p_\sigma t - q_\sigma) + (r_\sigma(\alpha, \beta) + p_\sigma\alpha)t + (p_\sigma\beta + q_\sigma + s_\sigma(\alpha, \beta)) \\ &= \chi_\sigma(\alpha, \beta)\tilde{\sigma}(t) + (r_\sigma(\alpha, \beta) + p_\sigma(\alpha - \chi_\sigma(\alpha, \beta)))t + (s\sigma(\alpha, \beta) + p_\sigma\beta + q_\sigma(1 - \chi_\sigma(\alpha, \beta))) \end{aligned} \tag{36}$$

The first part of the proposition now follows from uniqueness of the intrinsic character, linear, and constant residual and the fact that $\tilde{\sigma}$ is centered.

We prove the correctness of Table 1 in Appendix C.

□

Proposition 1.6. *If σ is uncoupling, then so is any affine shift $\varsigma(t) = \gamma\sigma(t) + \delta t + \epsilon$, for $\gamma \neq 0$.*

Proof. For $m, n \in \mathbb{Z}_+$, let $\{(w_i \in \mathbb{R}^m, b_i)\}_{i=1}^n$ be a family of non-degenerate and pairwise Γ_ς-separated pre-activations. Suppose that σ is uncoupling and let $\varsigma(t) = \gamma\sigma(t) + \delta t + \epsilon$, where $\gamma \neq 0$. By Lemma B.1, $\Gamma_\varsigma = \Gamma_\sigma$. Thus, for $a \in \mathbb{R}^n, c \in \mathbb{R}^m, d \in \mathbb{R}$, we have

$$\begin{aligned} 0 &= \sum_{i=1}^n a_i \varsigma(w_i \cdot x + b_i) + c \cdot x + d \\ &= \sum_{i=1}^n a_i \big(\gamma\sigma(w_i \cdot x + b_i) + \delta(w_i \cdot x + b_i) + \epsilon\big) + c \cdot x + d \\ &= \gamma \sum_{i=1}^n a_i \sigma(w_i \cdot x + b_i) + \big(c + \delta \sum_{i=1}^n a_i w_i\big) \cdot x + \big(d + \sum_{i=1}^n a_i(\delta b_i + \epsilon)\big), \quad \text{for every } x \in \mathbb{R}^m, \end{aligned}$$

if and only if

$$a = 0, \quad c = -\delta \sum_{i=1}^n a_i w_i, \quad d = -\sum_{i=1}^n a_i(\delta b_i + \epsilon),$$

since σ is uncoupling. However, this also forces $c = 0$ and $d = 0$, since $a = 0$. Thus, ς is uncoupling. □

Theorem 1.8. *Table 1 is correct and all activations in Table 1 are uncoupling.*

Proof. The correctness of Table 1 is shown for each activation separately in Appendix C. For the uncoupling property, the direction "⇐" of (10) is trivial. For "⇒", it suffices to show that $a = 0$, because then $c \cdot x + d = 0$ for all x and so $c = 0$ and $d = 0$. The proof goes via the following two lemmas that are proved after the proof of the current theorem is concluded.

Lemma B.2. *Let $\Gamma \leq \mathrm{Aff}_1$ consist of either only elements with $\alpha = 1$ or only elements with $\beta = 0$ (as is the case for all Γ_σ in Table 1). Let $\{(w_i, b_i)\}_{i=1}^n$ be a family of non-degenerate and pairwise Γ-separated pre-activations. Then we can find an affine line $t \mapsto v_0 + tv$ such that the (one-dimensional) pre-activations $\{(w_i \cdot v, w_i \cdot v_0 + b_i)\}_{i=1}^n$ are non-degenerate and pairwise Γ-separated.*

By applying Lemma B.2 we can restrict the left hand side of (10) to an affine line and obtain a family $\{(\omega_i, \tau_i)\}_{i=1}^n$ of non-degenerate and pairwise Γ_σ-separated one-dimensional pre-activations such that

$$\sum_{i=1}^n a_i \sigma(\omega_i t + \tau_i) + c't + d' = 0 \quad \text{for all } t \in \mathbb{R}, \tag{37}$$

where $\omega_i = w_i \cdot v$, $\tau_i = w_i \cdot v_0 + b_i$, $c' = c \cdot v$, $d' = c \cdot v_0 + d$. Thus, it remains to show that this implies that $a = 0$. We see in (37) that if we were able to find a linear "detector" functional L that is zero on affine functions and zero on all $\sigma(\omega_i t + \tau_i)$ except for some $i = i_0$, then we could isolate one term and infer $a_{i_0} = 0$. The idea of the next lemma is showing that this is possible, at least after potentially having massaged (37) a bit.

Lemma B.3. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function from Table 1 and let $\{(\omega_i, \tau_i)\}_{i=1}^n$ be a non-empty family of non-degenerate, pairwise Γ_σ-separated one-dimensional pre-activations. Then there exists $\Lambda \subset \mathbb{C}$, functions $\sigma_i : \Lambda \to \mathbb{C}$ for $i = 1, \ldots, n$, and an $\mathbb{R}$-linear functional $L : \mathrm{span}_{\mathbb{R}}(\sigma_1, \ldots, \sigma_n, z, 1) \to \mathbb{C}$, such that*

1. *whenever (37) holds, then*

$$\sum_{i=1}^n a_i \sigma_i(z) + c'z + d' = 0 \quad \text{for all } z \in \Lambda, \tag{38}$$

2. *L is zero on affine functions, and there exists an index i_0 such that $L\sigma_{i_0} \neq 0$, but $L\sigma_i = 0$ for $i \neq i_0$.*

The proof of Theorem 1.8 concludes by using Lemma B.3 to transfer (37) to (38), followed by applying L and inferring that $a_{i_0} = 0$ as explained earlier. Removing i_0 from the family of pre-activations and proceeding inductively we obtain that all $a_i = 0$. Therefore we have shown "$\Rightarrow$" in (10) and σ is uncoupling. □

In the proof of Lemma B.2 we use the following known result from linear algebra.

Proposition B.4. *Let V be an affine space over a field K with $|K| = k$. If for $\ell < k$ there exists a family $\{U_i\}_{i=1}^{\ell}$ of affine subspaces of V such that*

$$V = \bigcup_{i=1}^{\ell} U_i, \tag{39}$$

then there exists $j \in \{1, \ldots, \ell\}$ such that $U_j = V$.

Proposition B.4 has the following obvious corollary.

Corollary B.5. *A vector space over $\mathbb{R}$ is not the union of any finite number of proper affine subspaces.*

Proof of Lemma B.2. Let $\Gamma \leq \mathrm{Aff}_1$. For the non-degeneracy part, note that if for every line $t \mapsto v_0 + tv$, there exists i and $\alpha \in \mathbb{R} \setminus \{0\}$ such that $\alpha w_i \cdot v = 0$, then we must have $w_i = 0$ for some i by Corollary B.5. By contraposition the desired result follows. In the sequel, we can assume non-degeneracy.

For the pairwise Γ-separation, we divide the proof into two cases.

[Case 1]: ($\alpha = 1$, for every $(\alpha, \beta) \in \Gamma$)

With $\alpha = 1$, the action of $(\alpha, \beta) \in \Gamma$ on a pre-activation (w, b) is $(\alpha, \beta) \cdot (w, b) = (w, b + \beta)$. We argue by contraposition. Assume that for every line $t \mapsto v_0 + tv$, there exists $i \neq j$ and $(1, \beta) \in \Gamma$ such that

$$(w_i \cdot v, w_i \cdot v_0 + b_i + \beta) = (w_j \cdot v, w_j \cdot v_0 + b_j). \tag{40}$$

For every pair (i, j) satisfying (40) for some $t \mapsto v_0 + tv$, it holds that

$$w_i \cdot v = w_j \cdot v,$$

or equivalently

$$(w_i - w_j) \cdot v = 0.$$

Since there are only a finite number of pairs (i, j), there exists $i \neq j$ such that $w_i = w_j$ by Corollary B.5. In particular, there is at least one such pair $i \neq j$ satisfying (40). For such a pair the equation

$$w_i \cdot v_0 + b_i + \beta = w_j \cdot v_0 + b_j,$$

reduces to

$$b_i + \beta = b_j.$$

Thus, there exists $i \neq j$ and $(1, \beta) \in \Gamma$ such that

$$(w_i, b_i + \beta) = (w_j, b_j).$$

The desired implication follows by contraposition.

[Case 2]: ($\beta = 0$, for every $(\alpha, \beta) \in \Gamma$)

With $\beta = 0$, the action of $(\alpha, \beta) \in \Gamma$ on a pre-activation (w, b) is $(\alpha, \beta) \cdot (w, b) = (\alpha w, \alpha b)$. Let $\{(w_i, b_i)\}_{i=1}^{n}$ be a family of non-degenerate and pairwise Γ-separated pre-activations. For every $(\alpha, 0) \in \Gamma$ and every $i \neq j$, if $\alpha w_i = w_j$, then $\alpha b_i \neq b_j$ by pairwise Γ-separation, which implies that for any line $t \mapsto v_0 + tv$

$$\begin{aligned} \alpha w_i \cdot v_0 + \alpha b_i &= w_j \cdot v_0 + \alpha b_i \\ &\neq w_j \cdot v_0 + b_j. \end{aligned}$$

Which is what we wanted. Thus, we can assume $\alpha w_i \neq w_j$ for every $(\alpha, 0) \in \Gamma$ and every $i \neq j$. Now, fix an arbitrary v. If

$$\alpha w_i \cdot v \neq w_j \cdot v,$$

for every $i \neq j$ then we are done. We will therefore assume that there exists $i \neq j$ such that $\alpha w_i \cdot v = w_j \cdot v$. By non-degeneracy we can assume that $w_i \cdot v \neq 0$ for any i. Thus, we can write

$$\alpha = \frac{w_j \cdot v}{w_i \cdot v},$$

for such $i \neq j$. This means that for a fixed choice of v we have that $\alpha = \alpha_{ij}$ is determined by $i \neq j$. Because α is fixed, for $i \neq j$, the equations

$$(\alpha w_i - w_j) \cdot v_0 = b_j - \alpha b_i,$$

define at most $n(n-1)/2$ distinct affine subspaces. By Corollary B.5, there exists v_0 for which

$$(\alpha w_i - w_j) \cdot v_0 \neq b_j - \alpha b_i,$$

for every $i \neq j$. Which proves the result. □

Proof of Lemma B.3. For $\sigma = \mathrm{ReLU}$ we can use $\Lambda = \mathbb{R}$ and $\sigma_i(t) = \sigma(\omega_i t + \tau_i)$, which gives (38) immediately. Let i_0 be any index. Let d^+ and d^- denote the one-sided right and left derivative respectively and note that the one-sided derivatives are defined everywhere for σ_i. Write $k = -\tau_{i_0}/\omega_{i_0}$ and define

$$Lf = \mathrm{d}^+ f(k) - \mathrm{d}^- f(k). \tag{41}$$

L is zero on affine functions and on all σ_i with a different kink than k, since they are smooth at k. This covers all σ_i except σ_{i_0} due to Γ_σ-separatedness. Furthermore, $L\sigma_{i_0} = |\omega_{i_0}| \neq 0$.

For

$$\sigma(t) = \mathrm{ELU}_\lambda(t) = \begin{cases} t & t > 0 \\ \lambda\,(e^t - 1) & t \leq 0 \end{cases}, \tag{42}$$

we can use a variant of the idea used for ReLU. We again set $\Lambda = \mathbb{R}$ and $\sigma_i(t) = \sigma(\omega_i t + \tau_i)$. Let i_0 be any index. Define the polynomial $q(t) = t^2 \prod_{\omega_i \neq \omega_{i_0}} (t - \omega_i)$ and let $k = -\tau_{i_0}/\omega_{i_0}$. Now set

$$Lf = q(\mathrm{d}^+) f(k) - q(\mathrm{d}^-) f(k). \tag{43}$$

L is zero on affine functions due to the derivatives of degree ≥ 2 and engineered to select only σ_{i_0} as follows. If $\omega_i k + \tau_i \neq 0$, then σ_i is smooth at k, so the two derivatives are the same and cancel. If $\omega_i k + \tau_i = 0$, then one of the derivatives in (43) is on the affine side and one is on the exponential side of ELU. Compute for $m \geq 2$

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m} \sigma_i(t) = \begin{cases} 0 & \text{on the affine side,} \\ \lambda \omega_i^m e^{\omega_i t + \tau_i} & \text{on the exponential side.} \end{cases} \tag{44}$$

Therefore, when $\omega_i k + \tau_i = 0$, the affine side in (43) evaluates to 0 and the exponential side (by construction of q) to

$$\lambda q(\omega_i) \begin{cases} = 0 & \text{if } \omega_i \neq \omega_{i_0}, \\ \neq 0 & \text{if } \omega_i = \omega_{i_0}, \end{cases} \tag{45}$$

where the second case only happens for $i = i_0$ due to Γ_σ-separatedness.

For $\sigma = \mathrm{Sigmoid}$, we can use a complex "pole-hunting" argument as in Albertini et al. (1993) and Fefferman (1994). We set $\sigma_i(z) = \sigma(\omega_i z + \tau_i)$ with $\Lambda = \mathbb{C} \setminus \cup_i$ poles of σ_i. As Sigmoid is meromorphic on $\mathbb{C}$, we find that Λ only excludes a set of isolated points from $\mathbb{C}$, hence Λ is a domain. Further, Λ contains $\mathbb{R}$ and $0 \in \mathbb{R}$ is an accumulation point in Λ, so the identity theorem gives the transfer from (37) to (38). The poles of Sigmoid are at $(2k+1)\pi i$, $k \in \mathbb{Z}$. Take i_0 to be such that $|\omega_{i_0}|$ is maximal. Among all σ_i with the same $|\omega_i|$, the $\hat{\tau}_i = \mathrm{sign}(\omega_i)\tau_i$ must be distinct due to Γ_σ-separatedness. Thus, the poles of these σ_i are distinct. Let z_0 be one of the two poles of σ_{i_0} that are closest to the real axis. Since $|\omega_{i_0}|$ was maximal, z_0 can not be a pole of any σ_i with non-maximal $|\omega_i|$. Finally, we define

$$Lf = \lim_{z \to z_0} (z - z_0) f(z), \tag{46}$$

which is zero for all affine f, and zero for σ_i, $i \neq i_0$. $L\sigma_{i_0}$ is the residue of σ_{i_0} at the pole, namely $1/\omega_{i_0} \neq 0$. Similar pole-hunting arguments can be used for Tanh and SiLU.

For $\sigma = \mathrm{GELU}$ we can again use $\Lambda = \mathbb{R}$ and $\sigma_i(t) = \sigma(\omega_i t + \tau_i)$. However, we now need a good choice of i_0, in contrast to the ReLU and ELU cases (because GELU has no "kink"). We choose i_0 to be the index such that $(|\omega_{i_0}|, \mathrm{sign}(\omega_{i_0})\tau_{i_0})$ is smallest lexicographically. The index i_0 is unique due to Γ_σ-separatedness. Define

$$Lf = \lim_{t\to\infty} t^{-2} \exp\left((\omega_{i_0} t + \tau_{i_0})^2/2\right) \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(t). \tag{47}$$

The second derivative is zero on affine terms. Further, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{GELU}(\omega_i t + \tau_i) = \frac{\omega_i^2}{\sqrt{2\pi}}(2 - (\omega_i t + \tau_i)^2) \exp\left(-(\omega_i t + \tau_i)^2/2\right), \tag{48}$$

which lets us calculate $L\sigma_i$. When $|\omega_i| > |\omega_{i_0}|$ or when both $|\omega_i| = |\omega_{i_0}|$ and $\mathrm{sign}(\omega_i)\tau_i > \mathrm{sign}(\omega_{i_0})\tau_{i_0}$, then the exponential terms in (47) dominate and the limit is zero. If instead $i = i_0$, then the exponential terms cancel and the limit becomes the coefficient of t^2 in the numerator, which is $-\omega_{i_0}^4/\sqrt{2\pi} \neq 0$.

Finally, for $\sigma = \mathrm{Exp}$, we again use $\Lambda = \mathbb{R}$ and $\sigma_i(t) = \sigma(\omega_i t + \tau_i)$. Let i_0 be such that ω_{i_0} is the unique largest ω_i (possible due to Γ_σ-separatedness). Define

$$Lf = \lim_{t\to\infty} \exp(-\omega_{i_0} t)(f(t+2) - 2f(t+1) + f(t)). \tag{49}$$

The second difference is zero for affine functions. Further $L\sigma_i = 0$ for $i \neq i_0$ and $L\sigma_{i_0} = \exp(\tau_{i_0})(\exp(\omega_{i_0}) - 1)^2 \neq 0$.

□

The reason for not explicitly setting $\sigma_i(z) = \sigma(\omega_i z + \tau_i)$ in Lemma B.3 is to avoid ambiguities when extending to complex numbers and to enable more general constructions such as in the following proposition.

Proposition B.6. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be differentiable. If Lemma B.3 holds for $\varsigma(t) = \mathrm{d}/\mathrm{d}t\ \sigma(t)$ and $\Gamma_\varsigma \leq \Gamma_\sigma$, then Lemma B.3 holds for σ as well.*

Proof. Differentiate (37) once and in (38) use $\sigma_i = \omega_i \varsigma_i$, where ς_i are the functions that appear in (38) for ς, the lemma for ς gives $a = 0, c' = 0$ from which $d' = 0$ follows. □

Example B.7. For instance, since $\mathrm{Softplus}$ has derivative $\mathrm{Sigmoid}$ and both have Γ_σ containing only sign-flips, Proposition B.6 implies that $\mathrm{Softplus}$ is uncoupling.

B.2 Shallow Neural Networks

Proposition 1.22. *The network obtained by applying a network gauge defines the same function as the original network.*

Proof. From the definition of Γ_σ and the gauge transformation we have

$$\sigma(P_{(\gamma,\pi)}x + \beta) = K_{(\gamma,\pi)}\sigma(x) + R_{(\gamma,\pi)}x + s_\sigma(\gamma) \tag{50}$$

and

$$\begin{aligned} W' &= P_{(\gamma,\pi)}W, \qquad b' = P_{(\gamma,\pi)}b + \beta, \qquad A' = AK^{-1}_{(\gamma,\pi)}, \\ C' &= C - AK^{-1}_{(\gamma,\pi)}R_{(\gamma,\pi)}W, \qquad d' = d - AK^{-1}_{(\gamma,\pi)}(R_{(\gamma,\pi)}b + s_\sigma(\gamma)). \end{aligned} \tag{51}$$

Now, the evaluation of the transformed network is

$$\begin{aligned} &A'\sigma(W'x + b') + C'x + d' \\ &= AK^{-1}\sigma(PWx + Pb + \beta) + C'x + d' \\ &= AK^{-1}(K\sigma(Wx + b) + R(Wx + b) + s_\sigma(\gamma)) + C'x + d' \\ &= A\sigma(Wx + b) + AK^{-1}RWx + AK^{-1}(Rb + s_\sigma(\gamma)) + C'x + d' \\ &= A\sigma(Wx + b) + Cx + d. \end{aligned} \tag{52}$$

□

Proposition 1.23. *The gauge transformations form a group with composition law*

$$(\gamma_2, \pi_2)(\gamma_1, \pi_1) = (\gamma_2 \odot (\pi_2 \gamma_1), \pi_2 \pi_1), \tag{15}$$

where $\odot$ denotes coordinate-wise product in Γ_σ. The gauge transformation in (13) *is a left action of this group on network parameters.*

Proof. The identity gauge is given by $\pi = I$, α constant 1 and β constant 0. The inverse of a gauge is given by

$$(\gamma, \pi)^{-1} = (\pi^T \gamma^{\odot -1}, \pi^T). \tag{53}$$

It is straightforward to check the group axioms. The fact that we get a left action is also simply shown by tracing through definitions. □

Theorem 1.18. *Given a shallow neural network of width n with activation σ and a Γ_σ-normalizer η, there exists an η-canonical shallow neural network of width $n' \leq n$ computing the same function.*

Proof. If σ is affine, we can package the entire network as a width zero network by fusing $A\sigma(Wx + b)$ into $Cx + d$, which concludes the proof.

If σ is not affine, we proceed in five steps, transforming the given network into an η-canonical network. Let the given network be parameterized by (W, b, A, C, d). By deleting a neuron i, we mean removing the i'th column of A and the i'th row of W as well as the i'th element of b, thereby reducing the width of the network by one. Superscripts denote columns and subscripts rows.

1. Delete all neurons i such that A^i is zero.
2. Delete all neurons i such that W_i is zero. Compensate for each deletion by replacing d by $d + A^i \sigma(b_i)$.
3. Normalize all pre-activations by applying the network gauge with identity permutation and $\gamma_i = \eta(W_i, b_i)$.
4. Merge any equal pre-activations $i_1, i_2, \ldots i_k$ by replacing A^{i_1} with $\sum_{\ell=1}^{k} A^{i_\ell}$ and deleting the i_ℓ'th neurons for $\ell = 2, \ldots, k$. Also delete the new i_1'st neuron if the sum happens to be zero.
5. Sort all neurons so that the pre-activations are lexographically ordered. I.e. apply the network gauge with permutation that sorts the neurons and trivial γ.

After these five steps we obtain an η-canonical network and all steps preserve the computed function. □

Theorem 1.19. *Given an uncoupling activation σ and a Γ_σ-normalizer η, any two η-canonical shallow networks that define the same function must have the same parameters.*

Proof. Let the two η-canonical networks, of widths n and n' respectively, be parameterized by (W, b, A, C, d) and (W', b', A', C', d'). Look at any row j of the output, there the equality reads

$$\sum_{i=1}^{n} A_j^i \sigma(W_i x + b_i) + C_j x + d_j = \sum_{i=1}^{n'} A_j'^i \sigma(W_i' x + b_i') + C_j' x + d_j'. \tag{54}$$

Since all pre-activations in the two networks are non-degenerate and η-normalized, the set $S = \{(W_i, b_i)\}_{i=1}^{n} \cup \{(W_i', b_i')\}_{i=1}^{n'}$ contains non-degenerate and pairwise separated pre-activations. Fix some ordering on S and for $\iota = 1, \ldots, |S|$ define $E_j^\iota = A_j^{i(\iota)}$ if S_ι is a pre-activation $(W_{i(\iota)}, b_{i(\iota)})$ of the first network and $E_j^\iota = 0$ if S_ι is not a pre-activation of the first network. Similarly define $E_j'^\iota$ for the second network. From (54) we deduce that

$$\sum_{\iota=1}^{|S|} (E_j^\iota - E_j'^\iota) \sigma(W_\iota x + b_\iota) + (C_j - C_j') x + (d_j - d_j') = 0. \tag{55}$$

Since σ is uncoupling, it follows that $E^\iota_j - E'^\iota_j = 0, C_j - C'_j = 0, d_j - d'_j = 0$. We immediately obtain $C = C'$ and $d = d'$. Further, if a pre-activation S_ι is in S, then WLOG it is in the first network, but since the columns A^i are not zero, there must exist a j such that $A^{i(\iota)}_j \neq 0$, so that $E^\iota_j = E'^\iota_j$ forces S_ι to also exist in the second network. Therefore, the two networks contain the same pre-activations and since they are in the same lexicographical order we obtain $W = W'$, $b = b'$. Finally, $E^\iota_j = E'^\iota_j$ implies $A = A'$. □

Lemma B.8. *Let σ be an uncoupling activation and let an almost canonical shallow neural network with σ activation be given. If a gauge (γ, π) maps the network to itself, then the gauge is trivial (i.e. π is the identity and each $\gamma_i = (1, 0)$).*

Proof. Since the neurons of the network are pairwise Γ_σ-separated, it is impossible to have any non-trivial gauge such that

$$P_{(\gamma,\pi)}W = W \quad \text{and} \quad P_{(\gamma,\pi)}b + \beta = b. \tag{56}$$

□

Proposition 1.24. *Let σ be an uncoupling activation and assume that we are given two almost canonical shallow neural networks with σ activation that define the same function. Then the two networks have the same width and there exists a unique gauge transforming the first to the second.*

Proof. Since the networks are almost canonical, no step in the canonicalization procedure in the proof of Theorem 1.18 deletes any neuron. As the canonical form is unique, it follows that the two networks have the same canonical form, in particular the same width.

Let (γ, π) and (γ', π') be the gauges that the canonicalization procedure applies for each network. Since the canonical form is unique, the composed gauge $(\gamma', \pi')^{-1}(\gamma, \pi)$ maps the parameters of the first network to the parameters of the second network. If (γ'', π'') is any gauge that maps the parameters of the first network to the parameters of the second network, then $(\gamma, \pi)^{-1}(\gamma', \pi')(\gamma'', \pi'')$ maps the first network to itself, so it is the trivial gauge by Lemma B.8. It then follows that $(\gamma'', \pi'') = (\gamma', \pi')^{-1}(\gamma, \pi)$. □

B.3 Equivarianct Shallow Neural Networks

Theorem 2.1. *Let (W, b, A, C, d) be the parameters of an almost canonical shallow network with uncoupling activation σ. Assume that the network defines a function that is equivariant with respect to ρ_{in} and ρ_{out}. Then there exists a unique homomorphism $\Psi : G \to \Gamma_\sigma \wr S_n$ such that the following weight-tying equation holds:*

$$\mathcal{T}_{\Psi(g)}(W, b, \rho_{\text{out}}(g)A, \rho_{\text{out}}(g)C, \rho_{\text{out}}(g)d) = (W\rho_{\text{in}}(g), b, A, C\rho_{\text{in}}(g), d). \tag{17}$$

Proof. Equivariance of the network means

$$A\sigma(W\rho_{\text{in}}(g)x + b) + C\rho_{\text{in}}(g)x + d = \rho_{\text{out}}(g)A\sigma(Wx + b) + \rho_{\text{out}}(g)Cx + \rho_{\text{out}}(g)d. \tag{57}$$

This is clearly equivalent to the statement that for all $g \in G$, the two networks

$$\begin{aligned} O_g(W, b, A, C, d) &:= (W, b, \rho_{\text{out}}(g)A, \rho_{\text{out}}(g)C, \rho_{\text{out}}(g)d) \quad \text{and} \\ I_g(W, b, A, C, d) &:= (W\rho_{\text{in}}(g), b, A, C\rho_{\text{in}}(g), d) \end{aligned} \tag{58}$$

define the same function.

Since $\rho_{\text{in}}(g)$ and $\rho_{\text{out}}(g)$ are invertible, $W\rho_{\text{in}}(g)$ has non-zero rows whenever W has and $\rho_{\text{out}}(g)A$ has non-zero columns whenever A has. Furthermore, if $(\alpha, \beta) \in \Gamma_\sigma$ maps $(w_i\rho_{\text{in}}(g), b_i)$ to $(w_j\rho_{\text{in}}(g), b_j)$, then the same (α, β) also maps (w_i, b_i) to (w_j, b_j). Thus, Γ_σ-separatedness of the pre-activations in the original almost canonical network implies Γ_σ-separatedness of the pre-activations in the two transformed networks in (58). We conclude that the transformed networks are almost canonical.

Proposition 1.24 now tells us that for each $g \in G$, the pair of networks in (58) are related by a unique gauge (π_g, γ_g), mapping the first network to the second. Finally, we show that the map $g \mapsto (\gamma_g, \pi_g)$ is a group homomorphism. To this end, it suffices to show that $(\gamma_{g_2}, \pi_{g_2})(\gamma_{g_1}, \pi_{g_1})$ is a gauge from

$$(W, b, \rho_{\text{out}}(g_2g_1)A, \rho_{\text{out}}(g_2g_1)C, \rho_{\text{out}}(g_2g_1)d) \quad \text{to} \quad (W\rho_{\text{in}}(g_2g_1), b, A, C\rho_{\text{in}}(g_2g_1), d), \tag{59}$$

since then uniqueness of the gauge implies $(\gamma_{g_2}, \pi_{g_2})(\gamma_{g_1}, \pi_{g_1}) = (\gamma_{g_2 g_1}, \pi_{g_2 g_1})$.

We apply the gauge via the definition (13) on the first network. Since the gauge transformation is a left-action of the gauge group according to Proposition 1.23 and both O_g and I_g commute with applying a gauge, we can first apply $(\gamma_{g_1}, \pi_{g_1})$ to obtain

$$\begin{aligned} &\mathcal{T}_{(\gamma_{g_1}, \pi_{g_1})}(W, b, \rho_{\text{out}}(g_2 g_1)A, \rho_{\text{out}}(g_2 g_1)C, \rho_{\text{out}}(g_2 g_1)d) \\ &= (W\rho_{\text{in}}(g_1), b, \rho_{\text{out}}(g_2)A, \rho_{\text{out}}(g_2)C\rho_{\text{in}}(g_1), \rho_{\text{out}}(g_2)d) \end{aligned} \tag{60}$$

followed by applying (γ_{g_2}, pi_{g_2}) to obtain the second network in (59). □

Lemma B.9. *If two shallow neural networks with uncoupling activation σ are equivariant and their η-canonical forms have different homomorphisms Ψ and Ψ' satisfying* (17)*, then the two networks define different functions.*

Proof. Assume that the networks define the same function. Then they have the same canonical form, but then by Theorem 2.1 Ψ and Ψ' must be the same, which is a contradiction. □

Corollary 2.3. *Let two equivariant shallow almost canonical networks with uncoupling activation σ be given. Let Ψ and Ψ' be the unique homomorphisms specifying the weight-tying of the two networks, as in* (17)*. If Ψ and Ψ' are not conjugate, then the two networks define different functions.*

Proof. Denote the two networks by N and N'. For a proof by contradiction, assume that they define the same function. By Proposition 1.24 there is a unique gauge (γ, π) from N to N'. Application of a gauge commutes with applying I_g and O_g in (58). So by equivariance, both the gauges

$$\Psi'(g)(\gamma, \pi) \quad \text{and} \quad (\gamma, \pi)\Psi(g) \tag{61}$$

transform $O_g(N)$ to $I_g(N')$. Thus, by uniqueness of the gauge, we have $\Psi'(g) = (\gamma, \pi)\Psi(g)(\gamma, \pi)^{-1}$ which is a contradiction. □

B.4 Classification of Homomorphisms to the Gauge Group

Let K be a group and X a K-set. Write $K_x = \{k \in K \mid kx = x\}$ for the stabilizer of $x \in X$. Recall the following from elementary group theory (the orbit-stabilizer theorem, with corollaries).

Lemma B.10.

1. *For any K-set X and $x \in X$, we have an isomorphism of K-sets given by*
$$K/K_x \to Kx, \qquad kK_x \mapsto kx. \tag{62}$$
2. *Every nonempty transitive K-set is isomorphic to K/K' for some $K' \leq K$.*
3. *Two transitive K-sets K/K', K/K'' are isomorphic if and only if K' and K'' are conjugate in K.*
4. *Finite K-sets are classified, up to isomorphism, by finite multisets of conjugacy classes of finite-index subgroups of K. Taking representatives $K_1, \ldots, K_s$ of each conjugacy class from a multiset gives the K-set*
$$X = \bigsqcup_{i=1}^{s} K/K_i, \qquad |X| = \sum_{i=1}^{s} [K : K_i]. \tag{63}$$

The following theorem follows from Kovács' Uniqueness Theorem (Kovács, 1985; 1988) by decomposing the permutation part of each Ψ into orbits. We provide a full proof in our notation.

Proposition B.11. *Let Γ and G be groups. The conjugacy classes of homomorphisms $\Psi : G \to \Gamma \wr S_n$ are in bijection with multisets of equivalence classes of pairs (H_j, ϕ_j), where $H_j \leq G$, $\phi_j : H_j \to \Gamma$ are homomorphisms and $\sum_j [G : H_j] = n$. Two pairs (H, ϕ) and (H', ϕ') are considered equivalent if there exist $g \in G$, $q \in \Gamma$ such that $H' = gHg^{-1}$ and $\phi'(ghg^{-1}) = q\phi(h)q^{-1}$ for all $h \in H$.*

Proof. As a warmup, consider trivial Γ, so that we are looking at homomorphisms $\Psi : G \to S_n$. Thus the classification reduces to classifying all G-sets of size n up to conjugation by a permutation. But conjugating Ψ by a permutation $\pi \in S_n$ is the same thing as an isomorphism of G-sets. So we get the classification from Lemma B.10.

When Γ is not trivial, we again use orbit-stabilizer, but on the group $K = G \times \Gamma$. Given a homomorphism $\Psi : G \to \Gamma \wr S_n$, let $X_\Psi = [n] \times \Gamma$, which will be our underlying set. We now need to explain how K acts on X_Ψ and what this has to do with the statement to be proven.

Γ acts on X_Ψ on the right by

$$(i, a) \cdot c = (i, ac). \tag{64}$$

$(\gamma, \pi) \in \Gamma \wr S_n$ acts on X_Ψ on the left by

$$(\gamma, \pi) \cdot (i, a) = (\pi(i), \gamma_{\pi(i)} a). \tag{65}$$

The two given actions commute, since they are applied from different sides. Therefore, we get a *left*-action of $(\Gamma \wr S_n) \times \Gamma$ on X by

$$(\gamma, \pi, c) \cdot (i, a) = (\pi(i), \gamma_{\pi(i)} a c^{-1}). \tag{66}$$

$\Psi = (\gamma, \pi)$ now gives a K-action on X_Ψ by

$$(g, c) \cdot (i, a) = (\pi(g)(i), \gamma(g)_{\pi(g)(i)} a c^{-1}). \tag{67}$$

We first show that Ψ and Ψ' being equal up to conjugation corresponds exactly to isomorphism of the K-sets X_Ψ and $X_{\Psi'}$. Let $\Psi'(g)(b, p) = (b, p)\Psi(g)$ for a fixed gauge $(b, p) \in \Gamma \wr S_n$. We define $F : X_\Psi \to X_{\Psi'}$ by

$$F(i, a) = (b, p) \cdot (i, a) \tag{68}$$

F is K-equivariant because

$$F((\Psi(g), c) \cdot (i, a)) = (b, p)\Psi(g) \cdot (i, a) \cdot c^{-1} = \Psi'(g)(b, p) \cdot (i, a) \cdot c^{-1} = (\Psi'(g), c) \cdot F(i, a). \tag{69}$$

Thus F is an isomorphism of K-sets (with inverse given by acting by $(b, p)^{-1}$). In the other direction, assume an isomorphism $F : X_\Psi \to X_{\Psi'}$ is given. The orbits of the right Γ-action are of the form $\{i\} \times \Gamma$, and F must permute these orbits since it is equivariant under the Γ-action, so we can write

$$F(i, 1) = (p(i), b_{p(i)}). \tag{70}$$

However, this determines F completely:

$$F(i, a) = F((1, a^{-1}) \cdot (i, 1)) = (1, a^{-1}) \cdot F(i, 1) = (p(i), b_{p(i)} a) = (b, p) \cdot (i, a). \tag{71}$$

This means that Ψ and Ψ' are conjugate under (b, p), via (69).

It now remains to classify all K-sets X_Ψ stemming from a $\Psi : G \to \Gamma \wr S_n$.

Since Γ acts freely on the right, the orbits of X_Ψ under the K-action (67) are of the form $O \times \Gamma$, where O is an orbit of the permutation action π.

Looking at one such orbit $O \subset [n]$, we can pick an arbitrary base point $i_0 \in O$. Next, we define

$$H = \{h \in G \mid \pi(h) i_0 = i_0\}, \qquad \phi(h) = \gamma(h)_{i_0}. \tag{72}$$

Note that $|O| = [G : H]$ so H has finite index in G. That $\phi : H \to \Gamma$ is a homomorphism can be seen by writing out what $\Psi(hh') = \Psi(h)\Psi(h')$ means for $\gamma(h)_{i_0}$ when h stabilizes i_0. The stabilizer of $(i_0, 1) \in X_\Psi$ is now given by

$$S_{H,\phi} = \{(h, \phi(h)) \mid h \in H\}, \tag{73}$$

because $(g, c) \cdot (i_0, 1) = (i_0, 1)$ if and only if $\pi(g) i_0 = i_0$ and $\gamma(g)_{i_0} c^{-1} = 1$, i.e. $g \in H$ and $c = \phi(g)$.

By orbit-stabilizer, the K-set $O \times \Gamma$ is isomorphic to $K/S_{H,\phi}$ and two such K-sets are isomorphic precisely when their stabilizers $S_{H,\phi}$ and $S_{H',\phi'}$ are conjugate, i.e. when there exists $(g, c) \in K$ such that

$$(g, c) S_{H,\phi} (g, c)^{-1} = S_{H',\phi'}, \tag{74}$$

or equivalently

$$(g,q)(h,\phi(h))(g,q)^{-1} = (ghg^{-1}, q\phi(h)q^{-1}) \in S_{H',\phi'}, \tag{75}$$

which is precisely the equivalence between (H,ϕ) and (H',ϕ') described in the proposition. The conjugacy classes of homomorphisms Ψ hence correspond exactly to isomorphism classes of K-sets X_Ψ, and each such K-set can be split into orbits where each orbit is determined by an equivalence class of pairs (H,ϕ).

In the other direction, we need to show that any (H,ϕ) gives rise to a Ψ. This is done by the following construction. Choose coset representatives r_x for the cosets of H in G, and specifically choose $r_H = e$. Define $\pi(g)x = gx$ and

$$\gamma(g)_x = \phi(r_x^{-1} g r_{g^{-1}x}). \tag{76}$$

For $g, g' \in G$ and $x \in G/H$, we get

$$r_x^{-1} g r_{g^{-1}x} \in H, \qquad r_{g^{-1}x}^{-1} g' r_{g'^{-1}g^{-1}x} \in H, \tag{77}$$

and since ϕ is a homomorphism, we obtain

$$\begin{aligned} \gamma(gg')_x &= \phi(r_x^{-1} g g' r_{g'^{-1}g^{-1}x}) \\ &= \phi(r_x^{-1} g r_{g^{-1}x}) \phi(r_{g^{-1}x}^{-1} g' r_{g'^{-1}g^{-1}x}) \\ &= \gamma(g)_x \, \gamma(g')_{g^{-1}x}. \end{aligned} \tag{78}$$

Together with $\pi(gg') = \pi(g)\pi(g')$, we see that $\Psi = (\gamma, \pi)$ is a homomorphism. Furthermore, we have $\phi(h) = \gamma(h)_H$ by construction. Thus we get a homomorphism $\Psi : G \to \Gamma \wr S_{[G:H]}$ and combining multiple such gives a homomorphism $\Psi : G \to \Gamma \wr S_n$. □

Remark B.12. In Agrawal & Ostrowski (2022), G and consequently H is finite. Therefore any homomorphism $H \to \Gamma_{\mathrm{ReLU}}$ must map into $(\pm 1, 0)$. They state the above proposition in terms of H and the subgroup $K = \ker \phi$, which uniquely determines ϕ in that case.

C Proofs of Activation Properties

We go over the activations in Table 1 one by one and prove the correctness of that table.

Lemma C.1. *Let σ be an activation, let $u > 0$ be a constant and define $\hat{\sigma}(t) = \sigma(ut)$. If $(\alpha, 0) \in \Gamma_\sigma$, then $(\alpha, 0) \in \Gamma_{\hat{\sigma}}$, $\chi_{\hat{\sigma}}(\alpha, 0) = \chi_\sigma(\alpha, 0)$, $r_{\hat{\sigma}}(\alpha, 0) = r_\sigma(\alpha, 0)/u$, and $s_{\hat{\sigma}}(\alpha, 0) = s_\sigma(\alpha, 0)$. In particular, when $\beta = 0$ for all $(\alpha, \beta) \in \Gamma_\sigma$, we have that $\Gamma_\sigma = \Gamma_{\hat{\sigma}}$, $\chi_{\hat{\sigma}} = \chi_\sigma$,$r_{\hat{\sigma}} = u r_\sigma$, and $s_{\hat{\sigma}} = s_\sigma$.*

Proof. Let $(\alpha, 0) \in \Gamma_\sigma$. We calculate

$$\begin{aligned} \sigma(\alpha u t) &= \hat{\sigma}(\alpha t) \\ &= \chi_{\hat{\sigma}}(\alpha, 0)\hat{\sigma}(t) + r_{\hat{\sigma}}(\alpha, 0)t + s_{\hat{\sigma}}(\alpha, 0) \\ &= \chi_{\hat{\sigma}}(\alpha, 0)\sigma(ut) + r_{\hat{\sigma}}(\alpha, 0)t + s_{\hat{\sigma}}(\alpha, 0). \end{aligned} \tag{79}$$

The lemma follows by substituting $t' = ut$ in (79). □

C.1 ReLU

$$\mathrm{ReLU}(t) = \max(t, 0) \tag{80}$$

Assume that

$$\mathrm{ReLU}(\alpha t + \beta) = \gamma \mathrm{ReLU}(t) + \delta t + \epsilon. \tag{81}$$

Suppose $\beta > 0$. Then letting $t = 0$ in (81) yields $\epsilon = \beta$. In the case $\alpha > 0$, letting $t = -\beta/\alpha$ in (81) yields $\delta = \alpha$. On the other hand, letting $t = 1$ in (81) then yields $\gamma = 0$, by which it follows from (81) that $\mathrm{ReLU}(\alpha t + \beta) = \alpha t + \beta$, which is a contradiction. In the case $\alpha < 0$, letting $t = -1$ in (81) yields $\delta = \alpha$. On the other hand, letting $t = -\beta/\alpha$ in (81) yields $\gamma = 0$, by which it follows from (81) that $\mathrm{ReLU}(\alpha t + \beta) = \alpha t + \beta$ which is a contradiction.

Suppose now that $\beta < 0$. Then letting $t = 0$ in (81) yields $\epsilon = 0$. In the case $\alpha > 0$, letting $t = -\beta/\alpha$ yields $\delta = -\gamma$. On the other hand, letting $t = -1$ in (81) yields $\delta = 0$, which in turn yields $\gamma = 0$, by which it follows from $\mathrm{ReLU}(\alpha t + \beta) = 0$, which is a contradiction. In the case $\alpha < 0$, letting $t = 1$ in (81) yields $\delta = -\gamma$. On the other hand, letting $t = -\beta/\alpha$ in (81) yields $\delta = 0$, which in turn yields $\gamma = 0$, by which it follows from (81) that $\mathrm{ReLU}(\alpha t + \beta) = 0$.

Thus, $\beta = 0$ is the only possibility. Letting $t = 0$ in (81) then shows that $\epsilon = 0$. But then (81) reduces to

$$\mathrm{ReLU}(\alpha t) = \gamma \mathrm{ReLU}(t) + \delta t. \tag{82}$$

Suppose that $\alpha > 0$. Then we have that $\mathrm{ReLU}(\alpha t) = \alpha \mathrm{ReLU}(t)$, by which it follows from (82) that $\gamma = \alpha$ and $\delta = 0$. Now, suppose that $\alpha < 0$. Then we have that

$$\mathrm{ReLU}(\alpha t) = \mathrm{ReLU}(-|\alpha| t) = \max(-|\alpha| t, 0) = -\alpha \max(t, 0) + \alpha t = -\alpha \mathrm{ReLU}(t) + \alpha t, \tag{83}$$

by which it follows from (82) that $\gamma = -\alpha$ and $\delta = \alpha$.

Observe that the even part of ReLU can be written as

$$\mathrm{ReLU}(t) - \frac{1}{2} t, \tag{84}$$

and by Lemma B.1 it has the same intrinsic symmetry group as ReLU. For $\alpha > 0$ we have

$$\mathrm{ReLU}(\alpha t) - \frac{1}{2}\alpha t = \alpha \mathrm{ReLU}(t) - \frac{1}{2}\alpha t = \alpha(\mathrm{ReLU}(t) - \frac{1}{2} t). \tag{85}$$

For $\alpha < 0$ we have

$$\mathrm{ReLU}(\alpha t) - \frac{1}{2}\alpha t = (-\alpha)\mathrm{ReLU}(t) + \alpha t - \frac{1}{2}\alpha t = (-\alpha)(\mathrm{ReLU}(t) - \frac{1}{2} t). \tag{86}$$

So ReLU is centerable to its even part $\tilde{\sigma}(t) = \mathrm{ReLU}(t) - \frac{1}{2} t$. This proves that the contents of Table 1 hold for $\sigma = \mathrm{ReLU}$.

Now, two non-degenerate pre-activations (w, b) and (w', b') are in the same orbit under the action of Γ_{ReLU} if and only if $\alpha w = w'$ and $\alpha b = b'$ for some $\alpha \in \mathbb{R} \setminus \{0\}$. Let w_ι denote the first non-zero index of w, which exists for non-degenerate (w, b). Then $\eta(w, b) = (1/w_\iota, 0)$ is a Γ_{ReLU}-normalizer since it only maps pre-activations to each other if they are in the same orbit. For the normalized pre-activation (w^*, b^*), we clearly have $w^*_\iota = 1$. This proves that the contents of Table 3 hold for $\sigma = \mathrm{ReLU}$.

C.2 Sigmoid and Tanh

$$\mathrm{Sigmoid}(t) = \frac{1}{1 + e^{-t}} \qquad \mathrm{Tanh}(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}} = 2\,\mathrm{Sigmoid}(2t) - 1 \tag{87}$$

Assume that

$$\mathrm{Sigmoid}(\alpha t + \beta) = \gamma \mathrm{Sigmoid}(t) + \delta t + \epsilon. \tag{88}$$

Differentiating (88) once and using $\mathrm{Sigmoid}'(t) = \mathrm{Sigmoid}(t)(1 - \mathrm{Sigmoid}(t))$ yields

$$\alpha \mathrm{Sigmoid}(\alpha t + \beta)(1 - \mathrm{Sigmoid}(\alpha t + \beta)) = \gamma \mathrm{Sigmoid}(t)(1 - \mathrm{Sigmoid}(t)) + \delta. \tag{89}$$

Letting $t \to \infty$ in (89) yields $\delta = 0$.

In the case $\alpha < 0$, letting $t \to \infty$ yields $-\gamma = \epsilon$ and letting $t \to -\infty$ yields $\epsilon = 1$. Thus, $\gamma = -1$. Since $\mathrm{Sigmoid}(-t) = -\mathrm{Sigmoid}(t) + 1$, we must have $\alpha = -1$ and $\beta = 0$.

In the case $\alpha > 0$, letting $t \to \infty$ yields $\gamma = 1 - \epsilon$ and letting $t \to -\infty$ yields $\epsilon = 0$. Thus, $\gamma = 1$. Thus, we must have $\alpha = 1$ and $\beta = 0$.

Observe that the odd part of $\mathrm{Sigmoid}$ can be written as

$$\mathrm{Tanh}(t/2) = 2\mathrm{Sigmoid}(t) - 1, \tag{90}$$

and by Lemma B.1 it has the same intrinsic symmetry group as $\mathrm{Sigmoid}$. We only need to check $\alpha = -1$ for which we obviously have

$$\mathrm{Tanh}(\alpha t/2) = \alpha \mathrm{Tanh}(t/2), \tag{91}$$

since Tanh is odd. So $\mathrm{Sigmoid}$ is centerable to its odd part $\tilde{\sigma}(t) = \mathrm{Tanh}(t/2)$. This proves that the contents of Table 1 hold for $\sigma = \mathrm{Sigmoid}$. By Lemma C.1, it follows that the contents of Table 1 hold for $\sigma = \mathrm{Tanh}$.

Now, two non-degenerate pre-activations (w, b) and (w', b') are in the same orbit under the action of $\Gamma_{\mathrm{Sigmoid}} = \Gamma_{\mathrm{Tanh}}$ if and only if $(w, b) = (w', b')$ or $(-w, -b) = (w', b')$. Let w_ι denote the first non-zero index of w, which exists for non-degenerate (w, b). Then $\eta(w, b) = (\mathrm{sign}(w_\iota), 0)$ is a $\Gamma_{\mathrm{Tanh}} = \Gamma_{\mathrm{Sigmoid}}$-normalizer since it only maps pre-activations to each other if they are in the same orbit. For the normalized pre-activation (w^*, b^*), we clearly have $w^*_\iota > 0$. This proves that the contents of Table 3 hold for $\sigma = \mathrm{Sigmoid}$ and $\sigma = \mathrm{Tanh}$.

C.3 Softplus and SiLU

$$\mathrm{Softplus}(t) = \log(1 + e^t) \qquad \mathrm{SiLU}(t) = \frac{t}{1 + e^{-t}} = t\,\mathrm{Sigmoid}(t) \tag{92}$$

Assume that

$$\mathrm{Softplus}(\alpha t + \beta) = \gamma \mathrm{Softplus}(t) + \delta t + \epsilon. \tag{93}$$

Differentiating (93) once and using $\mathrm{Softplus}'(t) = \mathrm{Sigmoid}$ yields

$$\alpha \mathrm{Sigmoid}(\alpha t + \beta) = \gamma \mathrm{Sigmoid}(t) + \delta. \tag{94}$$

Dividing both sides of (94) by α and comparing with Table 1 for $\mathrm{Sigmoid}$ we see that $\alpha = \pm 1$, $\beta = 0$, $\gamma = 1$, $\delta = 0$ if $\alpha = 1$ and $\delta = -1$ if $\alpha = -1$. Letting $t = 0$ in (93) yields $\epsilon = 0$.

Observe that the even part of $\mathrm{Softplus}$ can be written as

$$\mathrm{Softplus}(t) - \frac{1}{2}t, \tag{95}$$

and by Lemma B.1 it has the same intrinsic symmetry group as $\mathrm{Softplus}$. For $\alpha = 1$ we have

$$\mathrm{Softplus}(\alpha t) - \frac{1}{2}\alpha t = \alpha \mathrm{Softplus}(t) - \frac{1}{2}\alpha t = \alpha(\mathrm{Softplus}(t) - \frac{1}{2}t). \tag{96}$$

For $\alpha = -1$ we have

$$\mathrm{Softplus}(\alpha t) - \frac{1}{2}\alpha t = (-\alpha)\mathrm{Softplus}(t) + \alpha t - \frac{1}{2}\alpha t = (-\alpha)(\mathrm{Softplus}(t) - \frac{1}{2}t). \tag{97}$$

So $\mathrm{Softplus}$ is centerable to its even part $\tilde{\sigma}(t) = \mathrm{Softplus}(t) - \frac{1}{2}t$. This proves that the contents of Table 1 hold for $\sigma = \mathrm{Softplus}$.

Now, assume that

$$\mathrm{SiLU}(\alpha t + \beta) = \gamma \mathrm{SiLU}(t) + \delta t + \epsilon. \tag{98}$$

Differentiating (98) once we get

$$\alpha((\alpha t + \beta)\mathrm{Sigmoid}'(\alpha t + \beta) + \mathrm{Sigmoid}(\alpha t + \beta)) = \gamma(t\mathrm{Sigmoid}'(t) + \mathrm{Sigmoid}(t)) + \delta \tag{99}$$

In the case $\alpha > 0$, letting $t \to -\infty$ in (99) yields $\delta = 0$. Letting $t \to \infty$ in (99) yields $\alpha = \gamma$. Letting $t = 0$ in (99) then yields

$$\beta \mathrm{Sigmoid}'(\beta) + \mathrm{Sigmoid}(\beta) = \frac{1}{2}, \tag{100}$$

which has solution $\beta = 0$. In the case $\alpha < 0$, letting $t \to -\infty$ in (99) yields $\alpha = \delta$. Letting $t \to \infty$ in (99) yields $0 = \gamma + \delta$ which means that $\gamma = -\alpha$. Letting $t = 0$ in again yields (100) which has the only solution $\beta = 0$.

Now, letting $t = 0$ in (98) yields $\epsilon = 0$. For $t \neq 0$, dividing both sides of (98) by t yields.

$$\alpha \mathrm{Sigmoid}(\alpha t) = \gamma \mathrm{Sigmoid}(t) + \delta. \tag{101}$$

Dividing both sides of (101) by α and comparing with Table 1 for $\mathrm{Sigmoid}$ we see that $\alpha = \pm 1$, $\gamma = 1$, $\delta = 0$ if $\alpha = 1$ and $\delta = -1$ if $\alpha = -1$.

Observe that the even part of SiLU can be written as

$$\mathrm{SiLU}(t) - \frac{1}{2}t, \tag{102}$$

and by Lemma B.1 it has the same intrinsic symmetry group as SiLU. For $\alpha = 1$ we have

$$\mathrm{SiLU}(\alpha t) - \frac{1}{2}\alpha t = \alpha \mathrm{SiLU}(t) - \frac{1}{2}\alpha t = \alpha(\mathrm{SiLU}(t) - \frac{1}{2}t). \tag{103}$$

For $\alpha = -1$ we have

$$\mathrm{SiLU}(\alpha t) - \frac{1}{2}\alpha t = (-\alpha)\mathrm{SiLU}(t) + \alpha t - \frac{1}{2}\alpha t = (-\alpha)(\mathrm{SiLU}(t) - \frac{1}{2}t). \tag{104}$$

So SiLU is centerable to its even part $\tilde{\sigma}(t) = \mathrm{SiLU}(t) - \frac{1}{2}t$. This proves that the contents of Table 1 hold for $\sigma = \mathrm{SiLU}$.

Now, two non-degenerate pre-activations (w, b) and (w', b') are in the same orbit under the action of $\Gamma_{\mathrm{Softplus}} = \Gamma_{\mathrm{SiLU}}$ if and only if $(w, b) = (w', b')$ or $(-w, -b) = (w', b')$. Let w_ι denote the first non-zero index of w, which exists for non-degenerate (w, b). Then $\eta(w, b) = (\mathrm{sign}(w_\iota), 0)$ is a $\Gamma_{\mathrm{SiLU}} = \Gamma_{\mathrm{Softplus}}$-normalizer since it only maps pre-activations to each other if they are in the same orbit. For the normalized pre-activation (w^*, b^*), we clearly have $w^*_\iota > 0$. This proves that the contents of Table 3 hold for $\sigma = \mathrm{Softplus}$ and $\sigma = \mathrm{SiLU}$.

C.4 GELU

$$\mathrm{GELU}(t) = t\Phi(t), \quad \Phi \text{ the standard normal CDF} \tag{105}$$

We have that if

$$\mathrm{GELU}(\alpha t + \beta) = \gamma \mathrm{GELU}(t) + \delta t + \epsilon, \tag{106}$$

then differentiating yields

$$\alpha\Phi(\alpha t + \beta) + \alpha^2 t\varphi(\alpha t + \beta) + \alpha\beta\varphi(\alpha t + \beta) = \gamma(\Phi(t) + t\varphi(t)) + \delta. \tag{107}$$

Differentiating once more yields

$$\alpha^2(2\varphi(\alpha t + \beta) + \alpha t\varphi'(\alpha t + \beta) + \beta\varphi'(\alpha t + \beta)) = \gamma(2\varphi(t) + t\varphi'(t)) \tag{108}$$

Consider the case $\alpha > 0$. Then letting $t \to -\infty$ in (107) yields $\delta = 0$ and letting $t \to \infty$ in (107) yields $\gamma + \delta = \alpha$ so $\gamma = \alpha$. Letting $t \to -\infty$ in (106) thus yields $\epsilon = 0$ and letting $t = 0$ in (106) then yields $\beta = 0$. Then letting $t = 0$ in (108) yields $\gamma = \alpha^2$, which means that $\alpha = 1$.

Consider the case $\alpha < 0$. Then letting $t \to \infty$ in (107) yields $\gamma + \delta = 0$ and letting $t \to -\infty$ in (107) yields $\delta = \alpha$, which yields $\gamma = -\alpha$. Letting $t \to \infty$ in (106) yields $\epsilon = 0$ and letting $t = 0$ in (106) then yields $\beta = 0$. Then letting $t = 0$ in (108) yields $\gamma = \alpha^2$, which means that $\alpha = -1$.

Observe that the even part of GELU can be written as

$$\mathrm{GELU}(t) - \frac{1}{2}t, \tag{109}$$

and by Lemma B.1 it has the same intrinsic symmetry group as GELU. For $\alpha = 1$ we have

$$\mathrm{GELU}(\alpha t) - \frac{1}{2}\alpha t = \alpha \mathrm{GELU}(t) - \frac{1}{2}\alpha t = \alpha(\mathrm{GELU}(t) - \frac{1}{2}t). \tag{110}$$

For $\alpha = -1$ we have

$$\mathrm{GELU}(\alpha t) - \frac{1}{2}\alpha t = (-\alpha)\mathrm{GELU}(t) + \alpha t - \frac{1}{2}\alpha t = (-\alpha)(\mathrm{GELU}(t) - \frac{1}{2}t). \tag{111}$$

So GELU is centerable to its even part $\tilde{\sigma}(t) = \mathrm{GELU}(t) - \frac{1}{2}t$. This proves that the contents of Table 1 hold for $\sigma = \mathrm{GELU}$.

Now, two non-degenerate pre-activations (w, b) and (w', b') are in the same orbit under the action of Γ_{GELU} if and only if $(w, b) = (w', b')$ or $(-w, -b) = (w', b')$. Let w_ι denote the first non-zero index of w, which exists for non-degenerate (w, b). Then $\eta(w, b) = (\mathrm{sign}(w_\iota), 0)$ is a Γ_{GELU}-normalizer since it only maps pre-activations to each other if they are in the same orbit. For the normalized pre-activation (w^*, b^*), we clearly have $w^*_\iota > 0$. This proves that the contents of Table 3 hold for $\sigma = \mathrm{GELU}$.

C.5 EXP

$$\mathrm{Exp}(t) = e^t \tag{112}$$

We have that if

$$\mathrm{Exp}(\alpha t + \beta) = \gamma \mathrm{Exp}(t) + \delta t + \epsilon, \tag{113}$$

then differentiating twice yields

$$\alpha^2 \mathrm{Exp}(\alpha t + \beta) = \gamma \mathrm{Exp}(t), \tag{114}$$

which is equivalent to

$$\alpha^2 \mathrm{Exp}((\alpha - 1)t)\mathrm{Exp}(\beta) = \gamma. \tag{115}$$

This means $\alpha = 1$, which gives $\gamma = e^\beta$, $\delta = 0$, and $\epsilon = 0$.

Note that since $\epsilon = \delta = 0$, we have that Exp is centered and thus p_σ and q_σ are trivial. This proves that the contents of Table 1 hold for $\sigma = \mathrm{Exp}$.

Now, two non-degenerate pre-activations (w, b) and (w', b') are in the same orbit under the action of Γ_{Exp} if and only if $w = w'$ and $b + \beta = b'$, for some $\beta \in \mathbb{R}$. Let w_ι denote the first non-zero index of w, which exists for non-degenerate (w, b). Then $\eta(w, b) = (1, -b)$ is a Γ_{Exp}-normalizer since it only maps pre-activations to each other if they are in the same orbit. For the normalized pre-activation (w^*, b^*), we clearly have $b^* = 0$. This proves that the contents of Table 3 hold for $\sigma = \mathrm{Exp}$.

C.6 ELU AND SELU

$$\mathrm{ELU}_\lambda(t) = \begin{cases} t & t > 0 \\ \lambda(e^t - 1) & t \leq 0 \end{cases} \qquad \mathrm{SELU}_{\kappa,\lambda}(t) = \kappa\, \mathrm{ELU}_\lambda(t) \qquad \text{with } \lambda \neq 0,\, \kappa \neq 0 \tag{116}$$

Assume that

$$\mathrm{ELU}_\lambda(\alpha t + \beta) = \gamma \mathrm{ELU}_\lambda(t) + \delta t + \epsilon. \tag{117}$$

Suppose that $\alpha < 0$. Then for $t > 0$, the right hand side of (117) is affine, but the left hand side decays exponentially as $t \to \infty$, which is a contradiction. Thus, we can assume that $\alpha > 0$.

Suppose that $\beta < 0$. Then for $t \in (0, -\beta/\alpha)$, the right hand side of (117) varies linearly in t, but the left hand side is exponential in t, which is a contradiction. Thus, we can assume that $\beta \geq 0$.

Suppose that $\beta \geq 0$. Letting $t = 0$ in (117) yields $\epsilon = \beta$. Letting $t = 1$ in (117) then yields $\delta = \alpha - \gamma$. For $t \in (-\infty, -\beta/\alpha)$ differentiating (117) once yields

$$\alpha\lambda e^{\alpha t + \beta} = \gamma \lambda e^t + \delta, \tag{118}$$

and differentiating (118) once more yields

$$\alpha^2 \lambda e^{\alpha t + \beta} = \gamma \lambda e^t = \alpha \lambda e^{\alpha t + \beta} - \delta. \tag{119}$$

Letting $t \to -\infty$ in (118) yields $\alpha - \gamma = \delta = 0$, and thus also $\gamma = \alpha = 1$ and $\epsilon = \beta = 0$, by (119).

Note that since $\delta = \epsilon = 0$, we have that ELU_λ is centered and thus p_σ and q_σ are trivial.

This proves that the contents of Table 1 hold for $\sigma = \mathrm{ELU}_\lambda$. By Lemma B.1 it follows that the contents of Table 1 hold for $\sigma = \mathrm{SELU}_{\kappa,\lambda}$.

Now, two non-degenerate pre-activations (w, b) and (w', b') are in the same orbit under the action of $\Gamma_{\mathrm{ELU}_\lambda} = \Gamma_{\mathrm{SELU}_{\kappa,\lambda}}$ if and only if $(w, b = (w', b')$. Let w_ι denote the first non-zero index of w, which exists for non-degenerate (w, b). Then $\eta(w, b) = (1, 0)$ is a $\Gamma_{\mathrm{ELU}_\lambda} = \Gamma_{\mathrm{SELU}_{\kappa,\lambda}}$-normalizer since it only maps pre-activations to each other if they are in the same orbit. Since the transformation $(1, 0)$ is trivial, any pre-activation (w^*, b^*) is normalized. This proves that the contents of Table 3 hold for $\sigma = \mathrm{ELU}_\lambda$ and $\mathrm{SELU}_{\kappa,\lambda}$.

D MINIMAL EXOTIC GELU ARCHITECTURE

Let the activation function

$$\sigma(x) = \mathrm{GELU}(x) := x\Phi(x) = x\frac{1}{2}(1 + \mathrm{erf}(\frac{x}{\sqrt{2}})), \tag{120}$$

where Φ is the standard normal CDF. Consider the shallow neural network $f_\theta : \mathbb{R} \to \mathbb{R}$, parametrized by $\theta = (W \in \mathbb{R}^{2\times 1}, A \in \mathbb{R}^{1\times 2})$,

$$f_\theta(x) = A\sigma(Wx). \tag{121}$$

If we have data (x, y) and a group $G = \{\pm 1\}$ acting on the input-output space by

$$\rho(g)(x, y) = (\rho(g)x, \mathrm{id}y) = \begin{cases} (x, y), & \text{if } g = 1 \\ (-x, y), & \text{if } g = -1. \end{cases} \tag{122}$$

A simple calculation, utilizing $\sigma(x) = \sigma(-x) + x$, shows that $f(-x) = f(x)$ for every $x \in \mathbb{R}$, if and only if A and W are picked such that $AW = 0$. The only choice of representation on the hidden layer that commutes with σ is to have $\rho(g) = P(g)$, where $P(g)$ is a permutation matrix (see Appendix E.2.1 in Godfrey et al. (2022)). The choice

$$A = \begin{pmatrix} b & b \end{pmatrix}, \qquad W = \begin{pmatrix} a \\ -a \end{pmatrix}, \quad a, b \in \mathbb{R},$$

is the classic layerwise equivariant parametrization, where the parameters are equivariant with respect to a permutation in the hidden layer. For $a \neq -b$, the choice of

$$A = \begin{pmatrix} -b & a \end{pmatrix}, \qquad W = \begin{pmatrix} a \\ b \end{pmatrix}, \quad a, b \in \mathbb{R},$$

corresponds to a representation on the hidden layer given by multiplication by -1, which does not commute with σ.

D.1 FUNCTION SPACE OF LAYERWISE AND EXOTIC GELU NETWORK

We get that the layerwise equivariant network is given by

$$f_\theta(x) = b(\sigma(ax) + \sigma(-ax)) = abx\frac{1}{2}\Big((\mathrm{erf}(\frac{ax}{\sqrt{2}}) - \mathrm{erf}(\frac{-ax}{\sqrt{2}})\Big) = abx\mathrm{erf}(\frac{ax}{\sqrt{2}}).$$

On the other hand, the exotic equivariant network is given by

$$f_\theta(x) = -b\sigma(ax) + a\sigma(bx) = \frac{abx}{2}\Big(\mathrm{erf}(\frac{bx}{\sqrt{2}}) - \mathrm{erf}(\frac{ax}{\sqrt{2}})\Big).$$

Here, the situation is less clear. If we pick a, b with opposite signs with $a < b$, then we have

$$f_\theta(x) = -\frac{|ab|x}{2}\Big(\mathrm{erf}(\frac{|b|x}{\sqrt{2}}) + \mathrm{erf}(\frac{|a|x}{\sqrt{2}})\Big)$$

and with $b < a$, we have

$$f_\theta(x) = -\frac{|ab|x}{2}\Big(-\mathrm{erf}(\frac{|b|x}{\sqrt{2}}) - \mathrm{erf}(\frac{|a|x}{\sqrt{2}})\Big) = \frac{|ab|x}{2}\Big(\mathrm{erf}(\frac{|b|x}{\sqrt{2}}) + \mathrm{erf}(\frac{|a|x}{\sqrt{2}})\Big).$$

In the limit when $|b| \to |a|$ we have that $f_\theta(x) \to -a^2x\mathrm{erf}(\frac{|a|x}{\sqrt{2}})$ if $a < b$ and $f_\theta(x) \to a^2x\mathrm{erf}(\frac{|a|x}{\sqrt{2}})$ when $b < a$. These are clearly achievable with the layer-wise equivariant parametrization. In the case $a < b$, we can choose $c = -|a|$ and $d = |a|$, which yields

$$-\frac{|ab|x}{2}\Big(\mathrm{erf}(\frac{|b|x}{\sqrt{2}}) + \mathrm{erf}(\frac{|a|x}{\sqrt{2}})\Big) - cdx\mathrm{erf}(\frac{dx}{\sqrt{2}}) = -\frac{|ab|x}{2}\Big(\mathrm{erf}(\frac{|b|x}{\sqrt{2}}) + \mathrm{erf}(\frac{|a|x}{\sqrt{2}})\Big) - (-a^2)x\mathrm{erf}(\frac{|a|x}{\sqrt{2}}) \to 0$$

as $|b| \to |a|$. The case $b < a$ is similar, except we choose $c = d = |a|$. We can also say something about what happens when $|a| \neq |b|$. In this case we can see that, since $\mathrm{erf}(x) \to \pm 1$ as $x \to \pm\infty$,

by choosing c and d such that $|cd| = |ab|$, both the layerwise equivariant network and the exotic equivariant network behave asymptotically like $|abx|$ as $|x| \to \infty$. How large x needs to be for the approximation to be reasonable will depend on the specific values of a and b. This behaviour explains the near zero values for large a, b in the second and fourth quadrants in Figure 3.

Now, what about the case when a and b have the same sign? Since f_θ is even, we will consider only the case when $x \geq 0$. For $x = 0$, we have that for the exotic equivariant network $f_\theta(x) = 0$, and if $0 < a < b$, then $\operatorname{erf}(\frac{bx}{\sqrt{2}}) - \operatorname{erf}(\frac{ax}{\sqrt{2}}) > 0$ and $\frac{abx}{2} > 0$ so that $f_\theta(x) > 0$. However, as $x \to \infty$, $f_\theta(x) \to 0$, since $\operatorname{erf}(\frac{bx}{\sqrt{2}}) - \operatorname{erf}(\frac{ax}{\sqrt{2}})$ decays much faster than x grows. For the layerwise equivariant network, we have already established that it behaves asymptotically like $c|dx|$ as $x \to \infty$.

D.2 Optimization Dynamics of Exotic Network

Let a smooth loss $(f_\theta(x), y) \mapsto \ell(f_\theta(x), y)$ be given. For the GELU network defined in (121), the forward pass is

$$z = Wx, \tag{123}$$
$$h = \sigma(z), \tag{124}$$
$$Y = A^T h, \tag{125}$$
$$L = \ell(Y, y), \tag{126}$$

We calculate the differentials

$$\mathrm{d}z = (\mathrm{d}W\, x, \mathrm{d}A)^T \tag{127}$$
$$\mathrm{d}h = (\sigma'(Wx)\mathrm{d}W\, x, \mathrm{d}A)^T \tag{128}$$
$$\mathrm{d}Y = (A\sigma'(Wx)\mathrm{d}W\, x, \sigma(Wx)^T \mathrm{d}A)^T \tag{129}$$
$$\mathrm{d}L = (\ell'(Y, y)A\sigma'(Wx)\mathrm{d}W\, x, \ell'(Y, y)\sigma(Wx)^T \mathrm{d}A)^T \tag{130}$$

Thus, for the group orbit data augmented loss $L_G = \mathbb{E}_G[L]$ the gradients with respect to W and A in $\{\theta : AW = 0\}$ are given by

$$\nabla_W L_G(\theta) = (\sigma'(Wx)^T - \sigma'(-Wx)^T)A^T \ell'(Y, y)^T x^T = (\sigma'(Wx)^T - \frac{1}{2}\mathrm{id})A^T \ell'(Y, y)^T x^T$$

and

$$\nabla_A L_G(\theta) = (\sigma(Wx) + \sigma(-Wx))\ell'(y, t)^T = (\sigma(Wx) - \frac{1}{2}Wx)\ell'(y, t)^T.$$

If the corresponding risk is denoted $R^{\mathrm{aug}}(\theta) = \mathbb{E}_{x,y}[L_G]$ and

$$\dot{W} = -\nabla_W R^{\mathrm{aug}}(\theta) \tag{131}$$
$$\dot{A} = -\nabla_A R^{\mathrm{aug}}(\theta), \tag{132}$$

then in $\{\theta : AW = 0\}$ we have

$$T(W, A) \coloneqq A\dot{W} + \dot{A}W \tag{133}$$
$$= -\mathbb{E}_{x,y}\left[\ell'(f_\theta(x), y)\Big(xA\big(\sigma'(Wx)^T A^T\big) - \frac{x}{2}\|A\|^2 + \sigma(Wx)^T W - \frac{x}{2}\|W\|^2\Big)\right]. \tag{134}$$

We have for the layerwise equivariant version that $A\dot{W} + \dot{A}W = 0$ and for the exotic equivariant version that

$$T(A, W) = -\mathbb{E}_{x,y}\left[\ell'(f_\theta(x), y)\left(\Big(\frac{a^2 + b^2}{a}\sigma(ax) + \frac{a^2 + b^2}{b}\sigma(bx) - (a^2 + b^2)x\Big) + ab^2x^2\varphi(ax) + a^2bx^2\varphi(bx)\right)\right], \tag{135}$$

where φ is the normal pdf.

Let the loss ℓ be MSE, the ground truth be $y = x^2$, and x be distributed as a Rademacher r.v., $P(x = 1) = P(x = -1) = 1/2$. Then we can calculate (135). First, by using $\sigma(t) = t\Phi(t)$ and $1 - \Phi(t) = \Phi(-t)$, we can write

$$\frac{a^2 + b^2}{a}\sigma(ax) + \frac{a^2 + b^2}{b}\sigma(bx) - (a^2 + b^2)x = x(a^2 + b^2)(\Phi(bx) - \Phi(-ax)). \tag{136}$$

By factoring we also have

$$ab^2x^2\varphi(ax) + a^2bx^2\varphi(bx) = abx^2(b\varphi(ax) + a\varphi(bx)). \tag{137}$$

Writing out ℓ' and f_θ we have

$$\ell'(f_\theta(x), y) = 2(abx(\Phi(bx) - \Phi(ax)) - y). \tag{138}$$

Rewriting (135) by (136), (137), and (138) and evaluating the expected value along $y = x^2$, using the fact that the expected value of a function of a Rademacher r.v. is the evaluation of the even part of the function at 1, yields

$$T(W, A) = -2\Big(ab\big(\Phi(b) - \Phi(a)\big) - 1\Big)\Big(ab\big(b\varphi(a) + a\varphi(b)\big) + (a^2 + b^2)\big(\Phi(b) - \Phi(-a)\big)\Big), \tag{139}$$

since the expression under the expectation is, in fact, even in x.