On Universality Classes of Equivariant Networks

Marco Pacini* **Gabriele Santin†** **Bruno Lepri‡** **Shubhendu Trivedi§**

Abstract

Equivariant neural networks provide a principled framework for incorporating symmetry into learning architectures and have been extensively analyzed through the lens of their *separation power*, that is, the ability to distinguish inputs modulo symmetry. This notion plays a central role in settings such as graph learning, where it is often formalized via the Weisfeiler–Leman hierarchy. In contrast, the *universality* of equivariant models—their capacity to approximate target functions—remains comparatively underexplored. In this work, we investigate the approximation power of equivariant neural networks beyond separation constraints. We show that separation power does not fully capture expressivity: models with identical separation power may differ in their approximation ability. To demonstrate this, we characterize the universality classes of shallow invariant networks, providing a general framework for understanding which functions these architectures can approximate. Since equivariant models reduce to invariant ones under projection, this analysis yields sufficient conditions under which shallow equivariant networks fail to be universal. Conversely, we identify settings where shallow models do achieve separation-constrained universality. These positive results, however, depend critically on structural properties of the symmetry group, such as the existence of adequate normal subgroups, which may not hold in important cases like permutation symmetry.

1 Introduction

Equivariant neural networks offer a principled framework to incorporate symmetry into learning architectures, attracting sustained attention for both their empirical successes and theoretical richness [1–5]. While their separation power—the capacity to distinguish inputs up to symmetry—has been extensively studied, comparatively less is understood about their approximation capabilities.

In classical approximation theory, expressivity is often characterized via universality—the capacity of a model class to approximate any target function within a given function space to arbitrary precision [6, 7]. In the equivariant setting, however, this notion must be refined. This is because such models treat symmetric inputs as indistinguishable; they can only approximate functions compatible with the underlying symmetry, subject additionally to spurious constraints arising from the imperfect interactions between equivariance and linear inductive biases on neural network layers. In this context, universality becomes inherently relative—defined with respect to a particular separation relation that circumscribes the model's ability to distinguish inputs.

Graph learning has served as a primary testbed for studying invariant and equivariant architectures, where models are typically required to respect node permutation symmetries [4, 8, 9]. Within this setting, separation power is most commonly assessed using the Weisfeiler-Leman (WL) test [10] or homomorphism counting techniques [11]. A whole range of architectures—including Graph

*University of Trento; Fondazione Bruno Kessler. mpacini@fbk.eu
†Ca' Foscari University of Venice. gabriele.santin@unive.it
‡Fondazione Bruno Kessler, lepri@fbk.eu
§shubhendu@csail.mit.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Neural Networks (GNNs) [12–14], Invariant Graph Networks (IGNs) [4, 15], and subgraph-based models [16, 9]—have been analyzed through this lens. More recently, investigations analyzing separation power have been extended beyond graph-structured data to broader classes of equivariant models [17, 18]. While much of the literature in geometric deep learning has centered on separation as the primary metric of expressivity, recent work [19, 20] has called for a more comprehensive view that includes approximation capabilities more generally. The role of equivariant layers in determining approximation power remains underexplored. Typically, these are composed to increase the model's separation capacity, while a universal component—such as a multilayer perceptron with adjustable width—is appended to approximate functions within the separation-constrained class. As a result, universality is achieved only relative to the distinctions introduced by the equivariant backbone. However, there is no general theory describing how equivariant layers themselves contribute to approximation.

To address this gap, we examine the approximation capabilities of equivariant neural networks beyond what is captured by separation constraints alone. For this purpose, it suffices to focus on invariant architectures, as the core phenomena extend to the equivariant setting. Indeed, projecting the output of an equivariant model onto the trivial representation yields an invariant network. Accordingly, our analysis of invariant networks provides insight into the approximation limits of a broad class of equivariant architectures. We begin by showing that invariant neural networks can be expressed as function that vanishes on certain differential operators (Section 5.1). This formulation allows us to derive sufficient conditions under which a shallow invariant network fails to be universal within the class of separation-constrained continuous functions (Section 5.2).

Our theory and analysis leads to three key insights. First, remarkably, we identify network families that possess identical separation power yet differ in their approximation capabilities—demonstrating that separation alone does not fully characterize expressivity. In particular, we show that shallow networks composed of commonly used equivariant layers—such as PointNets and CNNs with filter width 1—fail to be universal, despite matching the separation power of permutation-invariant continuous functions (Section 6.1). Second, this implies that the only two architectural choices that impact approximation power are depth and the type of hidden representations, the latter being strongly influenced by the structure of the symmetry group. Third, we show that a generalization of the results by [5] produces a broad family of shallow models that are universal within the separation-constrained function class (Section 6.2). However, these constructions fundamentally rely on the structure of the symmetry group. In particular, on the existence of normal subgroups of suitable size, a condition that is not always met, as is the case for key symmetry groups such as the permutation group.

We summarize the main contributions of this work as follows:

- We characterize the universality classes of shallow invariant networks (Theorem 13).
- We establish general *sufficient conditions* under which universality fails, even within function classes exhibiting maximal separation (Theorem 14 and Theorem 15).
- Leveraging these results, we construct explicit examples of invariant models that attain maximal separation yet fail to be universal, demonstrating that separation is not sufficient to guarantee universality (Proposition 16).
- We generalize the results by Ravanbakhsh [5] to a broader family of models (Theorem 18).

2 Related Work

Classical approximation theory for neural networks has established foundational results for shallow architectures with sigmoidal activations [6, 7, 21]. Necessary and sufficient conditions on activation functions were later given by Leshno et al. [22], and further refinements appear in Pinkus [23]. For general treatments of approximation theory in modern neural networks, we refer to [24, 25].

Moving beyond shallow networks, Yarotsky [26, 27] proved fundamental results on the approximation rates of deep neural networks, while Siegel [28] derived sharp bounds for deep ReLU networks. These results establish that deep networks are not only universal under mild assumptions but also more parameter-efficient than their shallow counterparts, approximating complex functions with significantly fewer parameters.

Equivariant neural networks offer a principled way to encode symmetry into learning architectures [1–3], with early applications across physics [29], chemistry [30], biology [31], and computer vision [32]. Beyond the foundational work of Yarotsky [33], universality in equivariant and invariant settings has been studied from multiple perspectives. A number of works [5, 34–36] establish universality for certain shallow equivariant networks using unconstrained hidden representations. Keriven and Peyré [37] extended this analysis to equivariant graph neural networks. These proofs rely on two main techniques. The first is the application of the Stone–Weierstrass theorem, or one of its variants, for instance via invariant polynomials [38, 8], to establish density results in spaces of continuous functions. The second is the use of a symmetrization operator, which enforces equivariance but causes the dimension of intermediate representations to grow exponentially. However, these approaches are often impractical: the Stone–Weierstrass theorem cannot be applied directly, since the network families of interest do not form function algebras, while symmetrization leads to prohibitive computational costs. Although canonicalization methods can improve the efficiency of models derived through symmetrization [39, 40], in many cases such techniques cannot be applied [41] or remain computationally inefficient.

A complementary line of work examines permutation-equivariant networks over multisets. Zaheer et al. [42], Qi et al. [43], Segol and Lipman [44] prove universality for such models under constrained hidden representations, but their results are restricted to architectures of depth three. As a result, the universality of truly shallow networks—those with depth two or less—remained unresolved.

In this work, we address this gap and show that certain shallow equivariant networks are *not* universal in the space of equivariant functions. This stands in contrast to the fully connected case, where universality holds generically, and highlights that depth can play a qualitatively different role in equivariant architectures, extending beyond parameter efficiency to approximation capacity itself.

To capture practical models within a theoretical framework, recent work has shifted toward studying universality *up to separation*. In permutation-equivariant networks, expressivity has been analyzed through the Weisfeiler–Leman (WL) hierarchy [45–49], with refinements based on homomorphism counts and subgraph-aware techniques [50, 51]. Joshi et al. [17] extended this approach to geometric domains, deriving depth-sensitive universality results under representation and orbit separation constraints. More generally, Pacini et al. [18] recently characterized the separation power of neural networks for arbitrary finite groups and permutation representations. While these works elucidate distinguishability, they do not fully account for approximation behavior.

The role of equivariant layers in approximation, *beyond* their contribution to separation, remains only partially understood. In practice, such layers are often composed to enhance separation power, followed by a universal component—typically an MLP—to approximate functions within the induced separation class. In the invariant case, this composition can yield universality. In the equivariant case, however, universality is not guaranteed and is only known to hold in specific instances [44]. Thus, the expressive power of the overall model remains fundamentally limited by the separation achieved by the equivariant stack. Yet equivariant layers may also contribute directly to approximation, and in some cases are known to suffice for universality within a fixed separation class [5].

Our work provides a detailed analysis of the universality classes of shallow equivariant networks. We show that equivariant layers are *not always sufficient* to guarantee universality up to separation, and that separation alone is *not a complete proxy* for approximation. We show explicit examples of models with identical separation power but differing approximation capacity. More broadly, we introduce general techniques for comparing the approximation power of equivariant models beyond separation, offering a more refined and complete understanding of expressivity in symmetry-constrained architectures.

3 Preliminaries

3.1 Groups and Equivariance

We are interested in functions that exhibit symmetry under specified transformations. Mathematically, such symmetries are described by groups: sets of transformations closed under composition, equipped with inverses and an identity element. While group theory offers a rigorous algebraic framework for analyzing symmetry, applying these ideas within neural networks requires their reformulation in linear-algebraic terms. This translation is achieved via representation theory, which associates

abstract group elements with matrix actions on vector spaces. For a brief overview, see Appendix A; for a more detailed treatment, see [52].

Our focus will be on permutation representations, which naturally arise when a group G acts on a finite set X. Let $\mathbb{R}^X$ denote the space of real-valued functions on X. For each $x \in X$, define $e_x \in \mathbb{R}^X$ as the function taking value 1 at x and 0 elsewhere. The set $\{e_x\}_{x \in X}$ forms a canonical basis for $\mathbb{R}^X$. A permutation representation of G on $V = \mathbb{R}^X$ is a linear action satisfying $g(e_x) = e_{gx}$ for all $g \in G$ and $x \in X$. If V and W are permutation representations of G, a map $\phi : V \to W$ is G-equivariant if $\phi(gv) = g\phi(v)$ for all $g \in G$ and $v \in V$. We denote by $\mathrm{Hom}(V, W)$ the space of linear maps from V to W, and by $\mathrm{Hom}_G(V, W)$ the subspace of G-equivariant linear maps. Similarly, let $\mathrm{Aff}(V, W)$ denote the space of affine maps from V to W, and $\mathrm{Aff}_G(V, W)$ the subspace of G-equivariant affine maps. The spaces $\mathrm{Hom}(V, W)$, $\mathrm{Aff}(V, W)$, and their equivariant counterparts are real vector spaces under pointwise addition and scalar multiplication. A result from Pacini et al. [53] shows that any map $f \in \mathrm{Aff}(V, W)$ admits a unique decomposition of the form $f = \tau_v \circ \phi$ for some $v \in W$ and $\phi \in \mathrm{Hom}(V, W)$, where $\tau_v(w) = w + v$. Such a map is G-equivariant iff ϕ is G-equivariant and $v \in W^G = \{v \in W \mid gv = v; \forall g \in G\}$, the fixed-point subspace of W. In particular, there is a linear morphism $\lambda : \mathrm{Aff}_G(V, W) \to \mathrm{Hom}_G(V, W)$ that projects an affine map to its linear part.

3.2 Equivariant Neural Networks

With all necessary definitions in place, we now introduce the notion of an equivariant neural network. Throughout this work, we consider networks that are equivariant under the action of a finite group, using arbitrary point-wise continuous activation functions, and with layers that transform according to permutation representations. We adopt the notation introduced by Pacini et al. [18].

Definition 1 (Point-wise Activation)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a nonlinear activation function, and let $\mathbb{R}^X$ denote a permutation representation of a group G. We define the corresponding *point-wise activation* $\tilde{\sigma} : \mathbb{R}^X \to \mathbb{R}^X$ by setting $\tilde{\sigma}\left(\sum_{x \in X} \alpha_x e_x\right) = \sum_{x \in X} \sigma(\alpha_x) e_x$. When no confusion arises, we will denote both σ and $\tilde{\sigma}$ by the same symbol.

Definition 2 (Neural Networks and Neural Spaces)**.** Let G be a group, and let $V_0, \ldots, V_d$ be permutation representations of G. For each $i = 1, \ldots, d$, let $M^i \subseteq \mathrm{Aff}_G(V_{i-1}, V_i)$ be a set of G-equivariant affine maps. For $d \geq 2$, the *neural space* associated with the layers $M^1, \ldots, M^d$ and a point-wise activation function σ is defined recursively by

$$\mathcal{N}_\sigma(M^1, \ldots, M^d) = \left\{ \phi^d \circ \tilde{\sigma} \circ \eta^{d-1} \,\middle|\, \phi^d \in M_d,\ \eta^{d-1} \in \mathcal{N}_\sigma(M^1, \ldots, M^{d-1}) \right\},$$

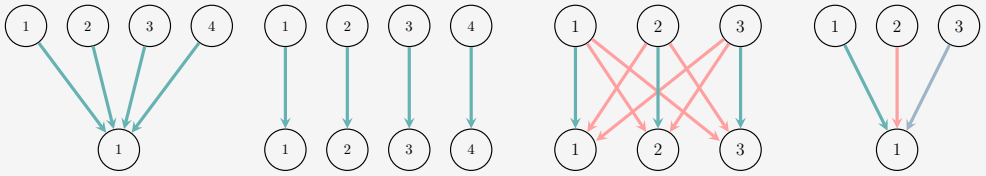
with the base case $\mathcal{N}_\sigma(M^1) = M^1$. An element $\eta^d \in \mathcal{N}_\sigma(M^1, \ldots, M^d)$ is called a *neural network* with layers in $M^1, \ldots, M^d$ and activation σ. When each M^i is taken to be the full space $\mathrm{Aff}_G(V_{i-1}, V_i)$, we write $\mathcal{N}_\sigma(V_0, \ldots, V_d)$ as shorthand for $\mathcal{N}_\sigma(M^1, \ldots, M^d)$.

To capture architectures commonly used in practice, we adopt a more structured form for the layer spaces $M \subseteq \mathrm{Aff}_G(V, \mathbb{R}^X)$, as proposed in Section 4.2 of Pacini et al. [18]. Specifically, we assume that M takes the form

$$M = \left\{ v \mapsto \sum_{i=1}^{k} x_i \phi^i(v) + \sum_{j=1}^{\ell} y_j \mathbb{1}_{X_j} \,\middle|\, x_1, \ldots, x_k, y_1, \ldots, y_\ell \in \mathbb{R} \right\}, \tag{1}$$

where $\phi^1, \ldots, \phi^k$ span a subspace of $\mathrm{Hom}_G(V, \mathbb{R}^X)$, $X_1, \ldots, X_\ell$ are the orbits of X under the G-action, and $\mathbb{1}_{X_i} := \sum_{x \in X_i} e_x$ for $i = 1, \ldots, \ell$. This formulation, while notation-heavy, plays a central role in the development of our main results.

We now present two working examples of equivariant affine maps and their associated neural spaces. These examples both reflect architectures commonly used in geometric deep learning and illustrate how standard models naturally conform to the structure in (1). They will serve as recurring reference points throughout to highlight key phenomena in the universality landscape of equivariant networks.



(a) $\mathrm{Aff}_{S_4}(\mathbb{R}^4, \mathbb{R})$ (b) C^1 (c) $\mathrm{Aff}_{S_4}(\mathbb{R}^4, \mathbb{R}^4)$ (d) $\mathrm{Aff}(\mathbb{R}^3, \mathbb{R})$

Figure 1: Colored bipartite graphs illustrating the layer spaces used in Examples 3 and 4, along with the graph corresponding to a standard fully connected layer, following the convention of Ravanbakhsh et al. [54] where arrows of the same color denote identical weights applied to different input values.

Example 3 (PointNets)**.** *We focus on the sum-pooling variant of PointNet architectures [43], which are designed to process unordered collections, such as point clouds, by enforcing permutation equivariance. An input configuration of n elements with f-dimensional features is represented by a tensor $A \in \mathbb{R}^{n \times f}$, where each row corresponds to the features of a single object. Permuting the elements corresponds to permuting the rows of A, i.e., the indices along its first axis. In our framework, the input tensor A is modeled as an element of $\mathbb{R}^X \otimes \mathbb{R}^f$, where $X = [n]$ and the symmetric group $G = S_n$ acts on X via its standard action and trivially on $\mathbb{R}^f$. PointNet architectures operate on such inputs using layers in the space $\mathrm{Aff}_{S_n}(\mathbb{R}^X \otimes \mathbb{R}^{f_{i-1}}, \mathbb{R}^X \otimes \mathbb{R}^{f_i})$, where each $\mathbb{R}^{f_i}$ corresponds to a space of S_n-invariant hidden features. Accordingly, the neural spaces corresponding to these equivariant architectures and their invariant counterparts take the following forms, respectively:*

$$\mathcal{N}_\sigma(\mathbb{R}^{X_0} \otimes \mathbb{R}^{f_0}, \ldots, \mathbb{R}^{X_d} \otimes \mathbb{R}^{f_d}) \quad \text{and} \quad \mathcal{N}_\sigma(\mathbb{R}^{X_0} \otimes \mathbb{R}^{f_0}, \ldots, \mathbb{R}^{X_{d-1}} \otimes \mathbb{R}^{f_{d-1}}, \mathbb{R}^{f_d}).$$

Zaheer et al. [42] showed that understanding the structure of $\mathrm{Aff}_{S_n}(\mathbb{R}^X \otimes \mathbb{R}^{f_{i-1}}, \mathbb{R}^X \otimes \mathbb{R}^{f_i})$ reduces to understanding $\mathrm{Aff}_{S_n}(\mathbb{R}^X, \mathbb{R}^X)$. Identifying $\mathbb{R}^X$ with $\mathbb{R}^n$, they established that

$$\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}^n) = \left\{ v \mapsto (x_1 \,\mathrm{id} + x_2 \, \mathbb{1}\mathbb{1}^\top)v + y\mathbb{1} \;\middle|\; x_1, x_2, y \in \mathbb{R} \right\},$$

where $\mathbb{1} = \mathbb{1}_{[n]} = [1, \ldots, 1]^\top$. Figure 1b shows the colored bipartite graph corresponding to the layer space $\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$. In the invariant case, $\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}) = \left\{ v \mapsto x \, \mathbb{1}^\top v + y \;\middle|\; x, y \in \mathbb{R} \right\}$, which is consistent with the notation introduced in (1)*. Figure 1a shows the colored bipartite graph corresponding to the layer space $\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R})$.*

Example 4 (Convolutional Neural Networks)**.** *Circular convolutional filters can be naturally formulated within the framework of permutation representations. For simplicity, we focus on the one-dimensional case. Let $X = [n]$ and let $G = \mathbb{Z}_n$ act on X by modular shifts. Identifying $\mathbb{R}^X$ with $\mathbb{R}^n$, the space $\mathrm{Hom}_{\mathbb{Z}_n}(\mathbb{R}^n, \mathbb{R}^n)$ corresponds to circulant matrices $A(x)$, each determined by a generating vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, as shown below.*

Each map in $\mathrm{Aff}_{\mathbb{Z}_n}(\mathbb{R}^n, \mathbb{R}^n)$ consists of a linear part defined by a circulant matrix and a bias term in $\mathbb{R}^n$:

$$A(x) := \begin{bmatrix} x_1 & x_n & x_{n-1} & \cdots & x_2 \\ x_2 & x_1 & x_n & \cdots & x_3 \\ x_3 & x_2 & x_1 & \cdots & x_4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & x_{n-1} & x_{n-2} & \cdots & x_1 \end{bmatrix} \quad \text{and } y\mathbb{1}_X = y\mathbb{1}_{[n]} = y \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Observe that any circulant matrix $A(x)$ can be written as a linear combination $A(e_1), \ldots, A(e_n)$, that is, $A(x) = \sum_{i=1}^n x_i A(e_i)$ where $\{e_1, \ldots, e_n\}$ denotes the standard basis of $\mathbb{R}^n$. Since limited-width convolutional filters are standard in practice, we restrict attention to the following maps:

$$C^k = \left\{ v \mapsto \sum_{i=1}^k x_i A(e_i)v + y\mathbb{1}_{[n]} \;\middle|\; x_1, \ldots, x_k, y \in \mathbb{R} \right\}. \tag{2}$$

This class can be seen as the one-dimensional analogue of the $k \times k$ convolutional kernels widely used in 2-D computer vision applications. The corresponding neural space is given by $\mathcal{N}_\sigma(C^{k_1}, \ldots, C^{k_d})$, for a choice of filter sizes $1 \leq k_1, \ldots, k_d \leq n$. Circular invariant layers can be characterized as

$$I := \mathrm{Aff}_{\mathbb{Z}_n}(\mathbb{R}^n, \mathbb{R}) = \left\{ v \mapsto (x \, \mathbb{1}^\top) \cdot v + y \;\middle|\; x, y \in \mathbb{R} \right\}. \tag{3}$$

In particular, we focus on the spaces C^1, which correspond to convolutional filters of width one; see Figure 1c for the colored bipartite graph representing the space $\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$.

Having detailed the structure of the layer spaces and their correspondence to practical architectures, we now turn to the study of universality in families of shallow neural spaces.

4 Universality in Shallow Neural Spaces

Universality Classes. To establish notation and introduce the notion of universality classes, we begin by reformulating the classical universality result for shallow neural networks [23] in terms of our framework. Observe that the full class of shallow neural networks with variable width can be written as $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(\mathbb{R}^m, \mathbb{R}^h, \mathbb{R})$. We denote by $\mathcal{U}_\sigma(\mathbb{R}^m, \mathbb{R}, \mathbb{R})$ the associated universality class—namely, the set of continuous functions on $\mathbb{R}^m$ approximable by such networks. Formally, $\mathcal{U}_\sigma(\mathbb{R}^m, \mathbb{R}, \mathbb{R})$ is defined as the closure of this union in $\mathcal{C}(\mathbb{R}^m)$, equipped with the topology of uniform convergence on compact sets.

Theorem 5. *The universality class for shallow neural networks, $\mathcal{U}_\sigma(\mathbb{R}^m, \mathbb{R}, \mathbb{R})$, coincides with $\mathcal{C}(\mathbb{R}^m)$ if and only if the activation function σ is not a polynomial.*

An analogous result in the equivariant setting was established by Ravanbakhsh [5] for neural networks defined on representations V and W as input and output spaces, respectively, and with regular hidden representations of the form $\mathbb{R}^G$. We define the universality class $\mathcal{U}_\sigma(V, \mathbb{R}^G, W)$ as the set of functions in $\mathcal{C}(V, W)$ that can be approximated by elements of $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(V, \mathbb{R}^G \otimes \mathbb{R}^h, W)$. Note that, in analogy with classical networks, the role of width is played by the hyperparameter h, which determines the dimension of the invariant hidden representation. The results of Ravanbakhsh [5] can then be stated as follows.

Theorem 6. *The universality class $\mathcal{U}_\sigma(V, \mathbb{R}^G, W)$ coincides with $\mathcal{C}_G(V, W)$, the space of continuous G-equivariant functions from V to W, if and only if the activation function σ is not a polynomial.*

We aim to provide a definition of universality classes that encompasses the notions introduced in Theorem 5 and Theorem 6, while being general enough to cover a wider range of architectures, such as PointNets and CNNs with variable filter size. To this end, we introduce the following auxiliary notation. Let G be a finite group and let V, W and Z be permutation representations of G. Let M be a subspace of $\mathrm{Aff}_G(V, W)$ and N a subspace of $\mathrm{Aff}_G(W, Z)$, as defined in (1). Then, for each $h \in \mathbb{N}$, we define M_h as the subspace of $\mathrm{Aff}_G(V, W \otimes \mathbb{R}^h)$ given by

$$M_h := \{x \mapsto (f_1(x), \dots, f_h(x)) \mid f_1, \dots, f_h \in M\}, \tag{4}$$

and define ${}_hN$ as the subspace of $\mathrm{Aff}_G(W \otimes \mathbb{R}^h, Z)$ given by

$${}_hN := \{(x_1, \dots, x_h) \mapsto g_1(x_1) + \cdots + g_h(x_h) \mid g_1, \dots, g_h \in N\}. \tag{5}$$

Recalling the isomorphism $W \otimes \mathbb{R}^h \cong (W)^{\oplus h}$, note that in the special cases where $M = \mathrm{Aff}_G(V, W)$ and $N = \mathrm{Aff}_G(W, Z)$, we have $M_h \cong \mathrm{Aff}_G(V, W \otimes \mathbb{R}^h)$ and ${}_hN \cong \mathrm{Aff}_G(W \otimes \mathbb{R}^h, Z)$. With this notation in place, we can now provide a general definition of universality classes.

Definition 7 (Universality Classes)**.** The universality class $\mathcal{U}_\sigma(M, N)$ associated with a family of neural spaces $\mathcal{N}_\sigma(M_h, {}_hN)$ for $h \in \mathbb{N}$ is the set of continuous functions approximated by these neural networks. More formally, $\mathcal{U}_\sigma(M, N)$ is defined as the closure of $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(M_h, {}_hN)$ in $\mathcal{C}(V, Z)$, equipped with the topology of uniform convergence on compact sets. As in the case of neural spaces, when $M = \mathrm{Aff}_G(V, W)$ and $N = \mathrm{Aff}_G(W, Z)$, we will simply write $\mathcal{U}_\sigma(V, W, Z)$.

However, comparing different universality classes is particularly challenging, and in the literature, separation power has often been used as a proxy for this purpose. The next section revisits this notion and critically examines its adequacy as a surrogate for universality.

On Separation-Constrained Universality. Theorem 6 establishes that equivariant neural networks cannot approximate all continuous functions. In particular, invariant networks are inherently unable to distinguish between symmetric inputs—a limitation that naturally constrains the class of functions they can represent. To make this precise, we formally define the notion of separation and the concept of *separation-constrained universality*.

Definition 8 (Separation-Constrained Universality)**.** A family of functions $\mathcal{N} \subseteq \{f : X \to Y\}$ is said to *separate* α and β if there exists $f \in \mathcal{N}$ such that $f(\alpha) \neq f(\beta)$. The set of point pairs not separated by $\mathcal{N}$ defines an equivalence relation:

$$\rho(\mathcal{N}) = \{(\alpha, \beta) \in X \times X \mid f(\alpha) = f(\beta) \text{ for all } f \in \mathcal{N}\}.$$

A family $\mathcal{N}$ is said to be *separation-constrained universal* if its relative universality class coincides with the entire set of continuous functions that respect the equivalence relation $\rho(\mathcal{N})$, that is,

$$\mathcal{C}_\rho(X, Y) = \{f \in \mathcal{C}(X, Y) \mid f(x) = f(y) \text{ for all } (x, y) \in \rho(\mathcal{N})\}.$$

It is a standard fact in approximation theory [55] that if a family of functions $\mathcal{N}$ fails to separate two points, then it cannot approximate any function that does. As such, separation-constrained universality captures the maximal expressivity achievable by $\mathcal{N}$. Here, we aim to investigate whether separation alone suffices to characterize expressivity i.e., whether universality classes with the same separation power must necessarily coincide. To this end, we now present three network families that share the same separation relation, despite differing in their internal representations. Throughout the remainder of the paper, we assume that all activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ are non-polynomial.

Proposition 9. [5] *Let C^1 be defined as in* (2)*, representing convolutional filters of width* 1*, and let I be as defined as in* (3)*, representing invariant circular layers. Let S_n act on $\mathbb{R}^n \cong \mathbb{R}^{[n]}$ via the standard permutation action. Then, the following universality classes have the same separation power:*

$$\rho\left(\mathcal{U}_\sigma(C^1, I)\right) = \rho\left(\mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^n, \mathbb{R})\right) = \rho\left(\mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^{S_n}, \mathbb{R})\right).$$

This naturally raises the following question.

Question 10. *Are these universality classes equal as well? More generally, is separation a complete proxy for comparing universality classes?*

We answer Question 10 in the negative via Proposition 16, after developing the necessary theory.

5 Main Results

In this section, we characterize the universality classes of invariant shallow neural networks (Section 5.1) and compare them (Section 5.2). Although the characterization is restricted to the invariant case, the following remark shows that it can be used to demonstrate that non-approximation in the invariant setting implies failure in the equivariant case as well.

Remark 11. Let $\mathcal{U}_1 \subseteq \mathcal{U}_2 \subseteq \mathcal{C}_G(V, W)$ be two universality classes with input space V and output space W. Let $\pi : W \to W^G$ denote the projection onto the trivial component of W. Define the pullback map

$$\pi^* : \begin{array}{ccc} \mathcal{C}_G(V, W) & \longrightarrow & \mathcal{C}_G(V, W^G) \\ f & \longmapsto & \pi \circ f, \end{array}$$

where $\mathcal{C}_G(V, W)$ denotes the space of continuous equivariant functions from V to W. Then, $\pi^*(\mathcal{U}_1) \subsetneq \pi^*(\mathcal{U}_2)$ implies $\mathcal{U}_1 \subsetneq \mathcal{U}_2$, since π^* is a continuous linear operator. This shows that a sufficient condition for strict inclusion between spaces of invariant networks also yields a sufficient condition for strict inclusion between the corresponding spaces of equivariant networks.

With this observation, we now restrict our attention to invariant networks without loss of generality.

5.1 Characterization of Universality Classes

To characterize the universality classes of invariant shallow networks, we begin by introducing the notion of a *basis map*.

Definition 12 (Basis maps)**.** As defined in (1), let M be a subspace of $\mathrm{Aff}_G(V, \mathbb{R}^Y)$, where V is a permutation representation and Y is a finite G-set of cardinality ℓ, which we identify with $[\ell]$. Let $\phi^1, \ldots, \phi^m$ be a basis for the linear part of M, and for each $i \in Y$, define the linear maps

$$\phi_i : \begin{array}{c} \mathbb{R}^X \to \mathbb{R}^m \\ x \mapsto (\phi_i^1(x), \ldots, \phi_i^m(x)). \end{array} \tag{6}$$

We refer to the maps $\phi_1, \ldots, \phi_\ell$ as the *basis maps* associated with M or its basis $\phi^1, \ldots, \phi^m$.

[5] For clarity of presentation, all proofs are deferred to the Appendix, with the exception of Proposition 16.

We now state the central characterization theorem for universality classes in terms of differential constraints on invariant functions.

Theorem 13. *Let M and N be, respectively, subspaces of $\mathrm{Aff}_G(V, W)$ and $\mathrm{Aff}_G(W, \mathbb{R})$. Let f be an invariant function, then $f \in \mathcal{U}_\sigma(M, N)$ if and only if $P(\partial_1, \dots, \partial_d)f = 0$ for every polynomial P that vanishes on the spaces spanned by the rows $\phi_i^1, \dots, \phi_i^m$ of each basis map $\phi_1, \dots, \phi_\ell$, see* (6)*.*

Here, we assume $d = \dim V$, and let $P(\partial_1, \dots, \partial_d)$ denote the constant-coefficient linear differential operator associated with the polynomial P. The derivatives ∂_i on V are interpreted in the distributional sense; see [56] for details.

Although Theorem 13 provides a complete characterization of the universality classes for arbitrary families of neural spaces, this generality may come at the cost of practicality. Indeed, computing the exact set of polynomials P can be particularly challenging, due to the combinatorial complexity arising from the intersections of the subspaces spanned by $\phi_i^1, \dots, \phi_i^m$. Nonetheless, the theorem is not merely of theoretical interest—it plays a central role in deriving sufficient conditions for universality failure. These conditions enable a principled comparison of the approximation power of distinct model families, as we explore in the following sections.

5.2 Sufficient Conditions for Universality Failure

In this section, we present two sufficient conditions for the failure of separation-constrained universality. These results will be used to resolve Question 10 and to prove Proposition 16. We begin with Theorem 14, which provides a general—but more difficult to verify—criterion, followed by Theorem 15, a less general version that is simpler to apply, despite its more convoluted appearance.

First, we introduce the notion of a directional derivative. For each vector $c = (c_1, \dots, c_n) \in \mathbb{R}^n$, the directional derivative is defined as the differential operator $D_c = c_1 \cdot \partial_1 + \cdots + c_n \cdot \partial_n$.

Theorem 14. *A continuous function f does not belong to the class $\mathcal{U}_\sigma(M, N)$ if*

$$D_{c_1} \cdots D_{c_\ell} f \neq 0 \tag{7}$$

for some choice of c_α in $\ker(\phi_\alpha^\top)$ for each basis map $\phi_1, \dots, \phi_\ell$.

In the case of equivariant networks where each affine layer is allowed to be an arbitrary equivariant affine map, Theorem 14 can be strengthened as follows.

Theorem 15. *Let $M = \mathrm{Aff}_G(V, W)$ and $N = \mathrm{Aff}_G(W, \mathbb{R})$, where V and W are permutation representations. Let $\phi_1, \dots, \phi_\ell$ denote the basis maps associated with M, see* (6)*. Then, the universal class $\mathcal{U}_\sigma(M, N)$ fails to be separation-constrained universal if, for some choice of:*

- *integers $s_1, \dots, s_\ell \in \{0, \dots, \ell\}$ satisfying $s_1 + \cdots + s_\ell = \ell$,*
- *integers $a_1 > \ell$ and $a_i + \ell < a_{i+1}$ for each $i = 1, \dots, \ell$,*
- *vectors $c_i \in \ker(\phi_i^\top)$ for each $i = 1, \dots, \ell$,*

Let $i_1, \dots, i_r$ be the indices such that $s_{i_j} \neq 0$. The following expression is nonzero:

$$\sum_{\sigma \in S_\ell} \frac{a_{i_1}!}{s_{i_1}!} \cdots \frac{a_{i_r}!}{s_{i_r}!} (c_{\sigma(1),1} \cdots c_{\sigma(s_1),1}) \cdot (c_{\sigma(s_1+1),2} \cdots c_{\sigma(s_1+s_2),2}) \cdots (c_{\sigma(\ell - s_\ell),\ell} \cdots c_{\ell,\ell}).$$

6 The Heterogeneous Landscape of Universality Classes

We now apply the tools developed in Section 5 to investigate the structure of universality classes and illustrate their heterogeneity. In Section 6.1, we address Question 10 by applying Theorems 14 and 15 to exhibit concrete examples of failure. In contrast, Section 6.2 presents Theorem 18, a generalization of Theorem 6, which provides sufficient conditions for achieving separation-constrained universality—highlighting the diversity of behaviors even within fixed symmetry classes.

6.1 Examples of Failure

Proposition 16. *As established in Proposition 9, the following spaces achieve the same separation power, yet differ in their approximation capabilities when $n > 2$:*

$$\mathcal{U}_\sigma(C^1, I) \subsetneq \mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^n, \mathbb{R}) \subsetneq \mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^{S_n}, \mathbb{R}).$$

By Remark 11, the corresponding equivariant models also have distinct approximation power.

We will prove the two strict inclusions of Proposition 16 in the following three paragraphs.

Failure for CNN with filter width 1: We now apply Theorem 14 to show that CNNs with filter width 1 cannot approximate the function $(x_1+\cdots+x_n)^n$ for $n > 1$, namely $(x_1+\cdots+x_n)^n \notin \mathcal{U}_\sigma(C^1, I)$. Indeed, for any $\alpha = 1, \ldots, n$, we have $e_{\alpha+1} \in \ker(\pi_\alpha^\top) = \mathrm{Span}\{e_1, \ldots, \hat{e}_\alpha, \ldots, e_n\}$, where $\alpha + 1$ is modulo n. Moreover, note that $D_{e_\alpha} = \partial_\alpha$, thus $\partial_n \cdots \partial_1 (x_1 + \cdots + x_n)^n = n! \neq 0$, which violates (7) in Theorem 14.

Success for PointNet: We now show that shallow PointNets approximate the polynomial function $(x_1+\cdots+x_n)^n$. By Proposition 41 in Appendix D, $f(x_1, x_1+\cdots+x_n)+\cdots+f(x_n, x_1+\cdots+x_n)$ belongs to $\mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^n, \mathbb{R})$ for any $f \in \mathcal{C}(\mathbb{R}^2)$. In particular, for $f(x, y) := y^n \in \mathcal{C}(\mathbb{R}^2)$, we see that $(x_1 + \cdots + x_n)^n \in \mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^n, \mathbb{R})$. Together with the previous observation, this establishes the first strict inclusion in Proposition 16, namely $\mathcal{U}_\sigma(C^1, I) \subsetneq \mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^n, \mathbb{R})$.

Failure for PointNet: We now aim to show that shallow PointNets cannot approximate the polynomial function $x_1 \cdots x_n$, which is S_n-invariant and therefore should, in principle, be approximable in a separation-constrained setting. We distinguish two cases: $n > 3$ and $n = 3$. Note that for $n = 2$, the symmetric group S_2 is abelian, and universality follows directly from Theorem 6.
We start considering $(n > 3)$. We again employ Theorem 14 to show that shallow invariant PointNets cannot approximate $x_1 \cdots x_n$, and hence neither CNNs with filter width 1. Indeed, note that the basis maps for $\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$ in this case are given by $\phi_\alpha(x_1, \ldots, x_n) = (x_\alpha, x_1 + \cdots + x_n)$. In matrix form, we write $\phi_\alpha = [e_\alpha, \mathbb{1}]^\top$. We define $K_\alpha := \ker\left(\phi_\alpha^\top\right) = \mathrm{Span}(e_i - e_j)_{i,j=1,\ldots,\hat{\alpha},\ldots,n}$. Then, define the following direction vectors:

$$\begin{aligned}
c_1 &:= e_2 - e_n \in K_1, & c_2 &:= e_3 - e_n \in K_2, \\
&\vdots \\
c_{n-3} &:= e_{n-2} - e_n \in K_{n-3}, & c_{n-2} &:= e_{n-1} - e_n \in K_{n-2}, \\
c_{n-1} &:= e_n - e_2 \in K_{n-1}, & c_n &:= e_1 - e_2 \in K_n.
\end{aligned}$$

Explicit computation shows that $D_{c_n} \cdots D_{c_1}(x_1 \cdots x_n) = 2$, verifying (7).
The previous technique does not apply in the case $n = 3$, for which we must instead resort to Theorem 15. First, define $c_1 := e_2 - e_3 \in K_1$, $c_2 := e_3 - e_1 \in K_2$, and $c_3 := e_1 - e_2 \in K_3$. Note that $c_{i,i} = 0$ for each $i = 1, 2, 3$. For $s_1 = 2$, $s_2 = 1$, and $s_3 = 0$, the polynomial becomes $a_1(a_1 - 1)a_2 \cdot [c_{3,1} \cdot c_{2,1} \cdot c_{1,2}] = -a_1(a_1 - 1)a_2 \neq 0$ by choosing $a_1, a_2 > 3$.

In view of the universality results for PointNet with depth 3 and arbitrary widths in both hidden layers by Segol and Lipman [44], this example highlights how, in the case of permutation equivariance, depth is crucial for achieving separation-constrained universality. This contrasts with other settings where universality can be achieved without relying on depth, as we will describe in the next section.

6.2 Examples of Separation-Constrained Universality

We now present Theorem 18, a generalization of Theorem 6, which shows that a specific class of hidden representations can achieve separation-constrained universality. These representations arise from cosets of particular subgroups H of G, defined as follows:

Definition 17 (Normal subgroup)**.** A subgroup H is *normal* if $ghg^{-1} \in H$ for each $h \in H, g \in G$.

Theorem 18. *Let V and Z be permutation representations of a finite group G, and let H be a normal subgroup of G. Therefore, $\mathcal{U}_\sigma(V, \mathbb{R}^{G/H}, Z)$ is separation-constrained universal.*

The converse does not always hold: representations arising from non-normal subgroups may nevertheless achieve separation-constrained universality, as illustrated in the following remark.

Remark 19. Let H be a non-normal subgroup of S_n contained in A_n. Then

$$\mathcal{U}_\sigma(\mathbb{R}^{S_n/A_n}, \mathbb{R}^{S_n/A_n}, \mathbb{R}) = \mathcal{U}_\sigma(\mathbb{R}^{S_n/A_n}, \mathbb{R}^{S_n/H}, \mathbb{R}) = \mathcal{C}_{S_n}(\mathbb{R}^{S_n/A_n}).$$

All subgroups of an abelian group are normal, whereas S_n has only one non-trivial normal subgroup, A_n, with $|S_n/A_n| = 2$, yielding hidden representations that are too small to be effective. We summarize by noting that intermediate representations built from abelian groups, such as those in standard circular CNNs, achieve separation-constrained universality. In contrast, architectures based on permutation representations lack this guarantee, as shown in Proposition 16.

7 Limitations

This work represents a first step toward understanding the approximation capabilities of equivariant networks beyond separation. Several limitations, however, remain. In particular, our analysis is limited to shallow networks. While these serve as minimal and analytically tractable examples, they may not fully capture the behavior of deeper architectures. Extending this framework to deeper networks—particularly in settings where depth interacts nontrivially with separation, as in IGNs—poses a significant challenge.

8 Conclusions

We investigated the approximation capabilities of equivariant neural networks, moving beyond their well-studied separation properties. By formulating shallow invariant networks as generalized superpositions of ridge functions (see Proposition 41), we developed a novel characterization of their universality classes and examined how architectural choices influence approximation behavior. Our analysis reveals that even networks with maximal separation power may fail to approximate all functions within the corresponding symmetry-respecting class, a phenomenon we attribute to the structure of their hidden representations. These findings suggest that approximation power cannot be deduced from separation alone and should be treated as a distinct axis of expressivity. Our results thus call for a more nuanced understanding of equivariant architectures—one that takes both axes into account in theoretical analysis and model design.

As future directions, we aim to extend our framework to determine whether failures of separation-constrained universality, such as those established in Proposition 16, persist in deeper architectures. Another important avenue for investigation is how differences in expressivity affect generalization, particularly among models that share the same separation power.

9 Acknoledgements

Bruno Lepri acknowledges the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and the support of the European Union's Horizon Europe research and innovation program under grant agreement No. 101120237 (ELIAS). This work was also partly supported by Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007) and European Union (Next Generation EU).

References

[1] Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2990–2999. PMLR, June 2016. URL https://proceedings.mlr.press/v48/cohenc16.html. ISSN: 1938-7228.

[2] Risi Kondor and Shubhendu Trivedi. On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2747–2755. PMLR, July 2018. URL https://proceedings.mlr.press/v80/kondor18a.html. ISSN: 2640-3498.

[3] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [cs, stat]*, May 2021. URL http://arxiv.org/abs/2104.13478. arXiv: 2104.13478.

[4] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and Equivariant Graph Networks. In *International Conference on Learning Representations*, September 2018. URL https://openreview.net/forum?id=Syx72jC9tm.

[5] Siamak Ravanbakhsh. Universal Equivariant Multilayer Perceptrons. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7996–8006. PMLR, November 2020. URL https://proceedings.mlr.press/v119/ravanbakhsh20a.html. ISSN: 2640-3498.

[6] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989. ISSN 1435-568X. doi: 10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274.

[7] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, January 1991. ISSN 0893-6080. doi: 10.1016/0893-6080(91)90009-T. URL https://www.sciencedirect.com/science/article/pii/089360809190009T.

[8] Omri Puny, Derek Lim, Bobak T. Kiani, Haggai Maron, and Yaron Lipman. Equivariant Polynomials for Graph Neural Networks, June 2023. URL http://arxiv.org/abs/2302.11556. arXiv:2302.11556 [cs].

[9] Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M. Bronstein, and Haggai Maron. Equivariant Subgraph Aggregation Networks, March 2022. URL http://arxiv.org/abs/2110.02910. arXiv:2110.02910 [cs, stat].

[10] B Yu Weisfeiler and A A Leman. THE REDUCTION OF A GRAPH TO CANONICAL FORM AND THE ALGEBRA WHICH APPEARS THEREIN. page 11, 1968.

[11] László Lovász. Large Networks and Graph Limits. volume 60 of *Colloquium Publications*, Providence, Rhode Island, December 2012. American Mathematical Society. doi: 10.1090/coll/060. URL http://www.ams.org/coll/060.

[12] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *Faculty of Informatics - Papers (Archive)*, January 2009. doi: 10.1109/TNN.2008.2005605. URL https://ro.uow.edu.au/infopapers/3165.

[13] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2, July 2005. doi: 10.1109/IJCNN.2005.1555942. ISSN: 2161-4407.

[14] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, February 2017. URL http://arxiv.org/abs/1609.02907. arXiv: 1609.02907.

[15] Haggai Maron, Or Litany, Gal Chechik, and Ethan Fetaya. On Learning Sets of Symmetric Elements. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6734–6744. PMLR, November 2020. URL https://proceedings.mlr.press/v119/maron20a.html. ISSN: 2640-3498.

[16] Emily Alsentzer, Samuel G. Finlayson, Michelle M. Li, and Marinka Zitnik. Subgraph Neural Networks, November 2020. URL http://arxiv.org/abs/2006.10538. arXiv:2006.10538 [cs, stat].

[17] Chaitanya K. Joshi, Cristian Bodnar, Simon V. Mathis, Taco Cohen, and Pietro Lio. On the Expressive Power of Geometric Graph Neural Networks. *International Conference of Learning Representations*, 2023. URL https://openreview.net/forum?id=Rkxj1GXn9_.

[18] Marco Pacini, Xiaowen Dong, Bruno Lepri, and Gabriele Santin. Separation Power of Equivariant Neural Networks, December 2024. URL http://arxiv.org/abs/2406.08966. arXiv:2406.08966 [cs].

[19] Christopher Morris, Fabrizio Frasca, Nadav Dym, Haggai Maron, İsmail İlkan Ceylan, Ron Levie, Derek Lim, Michael Bronstein, Martin Grohe, and Stefanie Jegelka. Future Directions in the Theory of Graph Machine Learning, June 2024. URL http://arxiv.org/abs/2402.02287. arXiv:2402.02287 [cs, stat].

[20] Yair Davidson and Nadav Dym. On the Hölder Stability of Multiset and Graph Neural Networks, April 2025. URL http://arxiv.org/abs/2406.06984. arXiv:2406.06984 [cs].

[21] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993. ISSN 1557-9654. doi: 10.1109/18.256500. URL https://ieeexplore.ieee.org/document/256500.

[22] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, January 1993. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80131-5. URL https://www.sciencedirect.com/science/article/pii/S0893608005801315.

[23] Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, January 1999. ISSN 1474-0508, 0962-4929. doi: 10.1017/S0962492900002919. URL https://www.cambridge.org/core/journals/acta-numerica/article/abs/approximation-theory-of-the-mlp-model-in-neural-networks/18072C558C8410C4F92A82BCC8FC8CF9.

[24] Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, May 2021. ISSN 0962-4929, 1474-0508. doi: 10.1017/S0962492921000052. URL https://www.cambridge.org/core/journals/acta-numerica/article/neural-network-approximation/7077A90FB36D405D903DCC82683B7A48.

[25] Philipp Petersen and Jakob Zech. Mathematical theory of deep learning, April 2025. URL http://arxiv.org/abs/2407.18384. arXiv:2407.18384 [cs].

[26] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, October 2017. ISSN 0893-6080. doi: 10.1016/j.neunet.2017.07.002. URL https://www.sciencedirect.com/science/article/pii/S0893608017301545.

[27] Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks, June 2018. URL http://arxiv.org/abs/1802.03620. arXiv:1802.03620 [cs].

[28] Jonathan W. Siegel. Optimal Approximation Rates for Deep ReLU Neural Networks on Sobolev and Besov Spaces, April 2024. URL http://arxiv.org/abs/2211.14400. arXiv:2211.14400 [cs, math, stat].

[29] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, May 2018. URL http://arxiv.org/abs/1802.08219. arXiv:1802.08219 [cs].

[30] Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet - a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, June 2018. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.5019779. URL http://arxiv.org/abs/1712.06113. arXiv:1712.06113 [physics].

[31] Chaitanya K. Joshi, Arian R. Jamasb, Ramon Viñas, Charles Harris, Simon V. Mathis, Alex Morehead, and Pietro Liò. gRNAde: Geometric Deep Learning for 3D RNA inverse design, April 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.03.31.587283.

[32] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D steerable CNNs: learning rotationally equivariant features in volumetric data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 10402–10413, Red Hook, NY, USA, 2018. Curran Associates Inc.

[33] Dmitry Yarotsky. Universal approximations of invariant maps by neural networks, April 2018. URL http://arxiv.org/abs/1804.10306. arXiv:1804.10306 [cs].

[34] Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, April 2020. ISSN 0002-9939, 1088-6826. doi: 10.1090/proc/14789. URL https://www.ams.org/proc/2020-148-04/S0002-9939-2019-14789-4/.

[35] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the Universality of Invariant Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4363–4371. PMLR, May 2019. URL https://proceedings.mlr.press/v97/maron19a.html. ISSN: 2640-3498.

[36] Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Universality of group convolutional neural networks based on ridgelet analysis on groups. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 38680–38694, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-71387-108-8.

[37] Nicolas Keriven and Gabriel Peyré. Universal Invariant and Equivariant Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/ea9268cb43f55d1d12380fb6ea5bf572-Abstract.html.

[38] Ben Blum-Smith and Soledad Villar. Machine learning and invariant theory. *Notices of the American Mathematical Society*, 70(08):1, September 2023. ISSN 0002-9920, 1088-9477. doi: 10.1090/noti2760. URL http://arxiv.org/abs/2209.14991. arXiv:2209.14991 [stat].

[39] Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame Averaging for Invariant and Equivariant Network Design. October 2021. URL https://openreview.net/forum?id=zIUyj55nXR.

[40] Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with Learned Canonicalization Functions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15546–15566. PMLR, July 2023. URL https://proceedings.mlr.press/v202/kaba23a.html. ISSN: 2640-3498.

[41] Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant Frames and the Impossibility of Continuous Canonicalization, June 2024. URL http://arxiv.org/abs/2402.16077. arXiv:2402.16077 [cs].

[42] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

[43] Charles R. Qi, Su, Hao, Mo, Kaichun, and Guibas, Leonidas J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.16. URL http://ieeexplore.ieee.org/document/8099499/.

[44] Nimrod Segol and Yaron Lipman. On Universal Equivariant Set Networks, January 2020. URL http://arxiv.org/abs/1910.02421. arXiv:1910.02421 [cs, stat].

[45] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks?, February 2019. URL http://arxiv.org/abs/1810.00826. Number: arXiv:1810.00826 arXiv:1810.00826 [cs, stat].

[46] Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4602–4609, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33014602. URL https://aaai.org/ojs/index.php/AAAI/article/view/4384.

[47] Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably Powerful Graph Networks. *International Conference of Learning Representations*, 2019.

[48] Zhengdao Chen, Soledad Villar, Lei Chen, and Joan Bruna. On the equivalence between graph isomorphism testing and function approximation with GNNs, May 2019. URL http://arxiv.org/abs/1905.12560. arXiv:1905.12560 [cs, stat].

[49] Waïss Azizian and Marc Lelarge. Expressive Power of Invariant and Equivariant Graph Neural Networks, June 2021. URL http://arxiv.org/abs/2006.15646. arXiv:2006.15646 [cs, stat].

[50] Bohang Zhang, Jingchu Gai, Yiheng Du, Qiwei Ye, Di He, and Liwei Wang. Beyond Weisfeiler-Lehman: A Quantitative Framework for GNN Expressiveness, January 2024. URL http://arxiv.org/abs/2401.08514. arXiv:2401.08514 [cs, math].

[51] Fabrizio Frasca, Beatrice Bevilacqua, Michael M. Bronstein, and Haggai Maron. Understanding and Extending Subgraph GNNs by Rethinking Their Symmetries, October 2022. URL http://arxiv.org/abs/2206.11140. arXiv:2206.11140 [cs].

[52] William Fulton and Joe Harris. *Representation Theory*, volume 129 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2004. ISBN 978-3-540-00539-1 978-1-4612-0979-9. doi: 10.1007/978-1-4612-0979-9. URL http://link.springer.com/10.1007/978-1-4612-0979-9.

[53] Marco Pacini, Xiaowen Dong, Bruno Lepri, and Gabriele Santin. A Characterization Theorem for Equivariant Networks with Point-wise Activations, January 2024. URL http://arxiv.org/abs/2401.09235. arXiv:2401.09235 [cs] version: 1.

[54] Siamak Ravanbakhsh, Jeff Schneider, and Barnabás Póczos. Equivariance Through Parameter-Sharing.

[55] DeVore. *Constructive Approximation*. 1993. URL https://link.springer.com/book/9783540506270.

[56] Grubb, G. *Distributions and Operators*, volume 252 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2009. ISBN 978-0-387-84894-5 978-0-387-84895-2. doi: 10.1007/978-0-387-84895-2. URL http://link.springer.com/10.1007/978-0-387-84895-2. ISSN: 0072-5285.

[57] M. F. Atiyah and I. G. MacDonald. *Introduction To Commutative Algebra*. Avalon Publishing, February 1994. ISBN 978-0-8133-4544-4. Google-Books-ID: HOASFid4x18C.

[58] Allan Pinkus. *Ridge Functions*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2015. ISBN 978-1-107-12439-4. doi: 10.1017/CBO9781316408124. URL https://www.cambridge.org/core/books/ridge-functions/25F7FDD1F852BE0F5D29171078BA5647.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The general investigation concerning the approximation power equivariant neural networks beyond separation constraints is described in Section 4. The main result, as mentioned in the abstract is described in section 5. The implications are described throughout the paper, but especially in section 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Separate limitation section is added to the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The definitions and notions are presented to facilitate a self-contained presentation of the main theorems. Additional details are relegated to the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA] .

   Justification: The paper presents a theoretical investigation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA] .

Justification: The paper presents a theoretical investigation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: The paper presents a theoretical investigation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: The paper presents a theoretical investigation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: The paper presents a theoretical investigation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All the authors have reviewed the code of ethics. The work presents theoretical results that analyze equivariant/invariant models. Most of the ethics concerns do not apply to such a work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: The work presents an analysis concerning a particular question regarding the universality and approximation power of equivariant models. As such it does not have broader social impacts that merit a separate discussion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: Theoretical paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: No experiments and data presented in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: No new assets introduced in the paper, given its theoretical nature.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: No crowdsourcing necessary for the presented work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The presented work does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: LLMs were not used to interrogate the results presented in the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Preliminaries

A.1 Group Theory

Definition 20 (Group)**.** A *group* is a set G equipped with a binary operation $\cdot : G \times G \to G$ satisfying the following properties:

- **Associativity:** for all $g, h, k \in G$, we have $(g \cdot h) \cdot k = g \cdot (h \cdot k)$.
- **Identity element:** there exists an element $e \in G$ such that $g \cdot e = e \cdot g = g$ for every $g \in G$.
- **Inverses:** for every $g \in G$, there exists an element $g^{-1} \in G$ such that $g \cdot g^{-1} = g^{-1} \cdot g = e$.

The group is said to be *finite* if G contains finitely many elements. It is called *abelian* (or *commutative*) if $g \cdot h = h \cdot g$ holds for all $g, h \in G$.

We now define the concept of a group homomorphism, a structure-preserving map between groups.

Definition 21 (Homomorphism)**.** Let G and H be groups. A function $\phi : G \to H$ is called a *group homomorphism* if, for all $g, h \in G$, it holds that

$$\phi(g \cdot h) = \phi(g) \cdot \phi(h).$$

Definition 22 (Cosets)**.** Let G be a group and let $H \leq G$ be a subgroup. The *left coset* of H associated with an element $g \in G$ is the set

$$gH = \{gh \mid h \in H\}.$$

The collection of all left cosets of H in G is denoted by $G/H = \{gH \mid g \in G\}$.

Similarly, the *right coset* of H corresponding to $g \in G$ is defined as

$$Hg = \{hg \mid h \in H\},$$

and the set of all left cosets is written as $G/H = \{gH \mid g \in G\}$.

Given another subgroup $K \leq G$, the *double coset* associated with $g \in G$ is the set

$$HgK = \{hgk \mid h \in H,\, k \in K\},$$

and the set of all such double cosets is denoted by $H\backslash G/K$.

Definition 23 (Normal subgroup)**.** A subgroup H is *normal* if $ghg^{-1} \in H$ for each $h \in H, g \in G$.

Example 24. *We highlight two families of normal subgroups relevant to our discussion:*

1. *All subgroups of abelian groups are normal.*
2. *The alternating group A_n is the only non-trivial normal subgroup of S_n.*

Theorem 25. *If H is a normal subgroup of G, then the cosets G/H for a group, where the binary operation is defined as $g_1H \cdot g_2H = g_1g_2H$.*

A.2 Group Actions and Equivariance

Let G be a group and let X be a set. A *group action* of G on X is a map

$$\Phi : G \times X \to X,$$

commonly written as $\phi_g(x) = \Phi(g, x)$ for $g \in G$ and $x \in X$, that satisfies the following two conditions:

- **Identity:** $\phi_e = \mathrm{id}_X$, where e is the identity element in G.
- **Compatibility:** For all $g, h \in G$, we have $\phi_g \circ \phi_h = \phi_{gh}$.

In practice, we often denote the action by $g \cdot x$ or simply gx in place of $\phi_g(x)$.

A set X endowed with a group action of G is referred to as a *G-set*. That is, X is a G-set if there exists a well-defined action $\cdot : G \times X \to X$ satisfying the identity and compatibility conditions above.

Another fundamental notion for our analysis is that of a map between G-sets that respects the group action. This leads to the definition of equivariance.

Definition 26 (Equivariance)**.** Let X and Y be G-sets. A function $f : X \to Y$ is said to be *G-equivariant* if, for all $g \in G$ and $x \in X$, the following condition holds:

$$f(g \cdot x) = g \cdot f(x).$$

A.3 Group Representations and Equivariant Affine Transformations

Let G be a group and V be a vector space over a field $\mathbb{R}$. A G-action $\Phi : G \times V \to V$ on V is *G-representation* if ϕ_g is linear for each g in G. Or equivalently,

$$\phi : \begin{array}{c} G \to \mathrm{GL}(V) \\ g \mapsto \phi_g \end{array}$$

where $\mathrm{GL}(V)$ is the general linear group of V, consisting of all invertible linear transformations on V. We will usually identify the entire $\Phi : G \times V \to V$ action with V itself and write $gv = \Phi(g, v)$.

Let V and W be two G-representations, we will indicate the set of equivariant linear maps between V and W as $\mathrm{Hom}_G(V, W)$ and as $\mathrm{Aff}_G(V, W)$ the set of equivariant affine maps. Note that $\mathrm{Hom}_G(V, W)$ is a vector space. Indeed, $0 \in \mathrm{Hom}_G(V, W)$ and for each $f, g \in \mathrm{Hom}_G(V, W)$ and each $\alpha, \beta \in \mathbb{R}$, $\alpha f + \beta g \in \mathrm{Hom}_G(V, W)$. The same is true for $\mathrm{Aff}_G(V, W)$.

Let V be a G-representation, we define the set of invariant vectors $V^G = \{v \in V \mid gv = v \ \forall g \in G\}$.

A.4 On Permutation Representations

Definition 27. Let X be a finite set and let G be a finite group acting on X. A *permutation representation* of G is the linear action of G on the space $\mathbb{R}^X$ defined by

$$g(e_x) = e_{g \cdot x} \quad \text{for all } g \in G,\ x \in X,$$

where $\{e_x\}_{x \in X}$ denotes the standard basis of $\mathbb{R}^X$.

Proposition 28. *Let X and Y be G-sets. Then, the following G-equivariant isomorphisms of representations hold:*

$$\mathbb{R}^{X \sqcup Y} \cong \mathbb{R}^X \oplus \mathbb{R}^Y \quad \textit{and} \quad \mathbb{R}^{X \times Y} \cong \mathbb{R}^X \otimes \mathbb{R}^Y,$$

where $X \sqcup Y$ denotes the disjoint union and $X \times Y$ the Cartesian product of the two sets.

B On Commutative Algebra

For a general introduction to commutative algebra, we refer to Atiyah and MacDonald [57]. Here, we recall the notation necessary to prove Theorem 13, 14 and 15.

Let $\mathbb{R}[x_1, \ldots, x_n]$ denote the set of polynomials in the variables $x_1, \ldots, x_n$.

Definition 29 (Ideal)**.** An *ideal* I of $\mathbb{R}[x_1, \ldots, x_n]$ is a subset such that, if $f \in I$, then $p \cdot f \in I$ for every $p \in \mathbb{R}[x_1, \ldots, x_n]$. If $X \subseteq \mathbb{R}^n$, we define

$$\mathcal{I}(X) = \{f \in \mathbb{R}[x_1, \ldots, x_n] \mid f(x) = 0 \ \forall x \in X\}.$$

Definition 30 (Product of Ideals)**.** Let $I, J \subseteq \mathbb{R}[x_1, \ldots, x_n]$ be ideals. Their product $I \cdot J$, or simply IJ, is the ideal defined by

$$IJ = \left\{ \sum_{k=1}^{r} f_k g_k \mid f_k \in I,\ g_k \in J,\ r \in \mathbb{N} \right\}.$$

Definition 31 (Generators of an Ideal)**.** Let $R = \mathbb{R}[x_1, \ldots, x_n]$ be the set of polynomial and let $f_1, \ldots, f_m \in R$. The *ideal generated* by $f_1, \ldots, f_m$ is the set

$$(f_1, \ldots, f_m) = \left\{ \sum_{i=1}^{m} h_i f_i \mid h_i \in R \right\}.$$

We say that $f_1, \ldots, f_m$ are *generators* of the ideal.

Proposition 32. *If X is a linear subspace of $\mathbb{R}^n$ such that its orthogonal complement $X^\perp$ is spanned by vectors $v_1, \ldots, v_d$, then*

$$\mathcal{I}(X) = (v_1^\top \cdot x, \ldots, v_d^\top \cdot x).$$

Proof. Indeed,

$$\mathcal{I}(X) \supseteq (v_1^\top \cdot x, \ldots, v_d^\top \cdot x).$$

To prove the reverse inclusion, observe that—up to a change of coordinates—we may assume $v_i \cdot x = x_i$ for $i = 1, \ldots, d$. In this case, any polynomial $f(x) \in \mathcal{I}(X)$ can be written as

$$f(x) = a_1(x)x_1 + \cdots + a_d(x)x_d + b(x),$$

where $a_i(x) \in \mathbb{R}[x_1, \ldots, x_n]$ for each $i = 1, \ldots, d$, and $b(x)$ is a polynomial whose monomials do not involve the variables $x_1, \ldots, x_d$.

Now, since f vanishes on $X = \{x \in \mathbb{R}^n : x_1 = \cdots = x_d = 0\}$, it must be that $b(x) = 0$ identically. Therefore, $f(x)$ lies in the ideal generated by $x_1, \ldots, x_d$, completing the proof. □

Remark 33. The following are either standard results or direct consequences of the observations above:

- The intersection and the product of ideals are themselves ideals.
- $\mathcal{I}(X_1 \cup \cdots \cup X_\ell) = \mathcal{I}(X_1) \cap \cdots \cap \mathcal{I}(X_\ell)$.
- If $X_1, \ldots, X_\ell$ are linear subspaces of $\mathbb{R}^n$, then $\mathcal{I}(X_1) \cdots \mathcal{I}(X_\ell)$ is generated by polynomials of the form $(v_1^\top \cdot x) \cdots (v_\ell^\top \cdot x)$, where $v_1, \ldots, v_\ell$ are vectors respectively in $X_1^\perp, \ldots, X_\ell^\perp$.

C On Superpositions of Ridge Functions

In this section, we present results on the theory of superpositions of generalized ridge functions. A detailed exposition can be found in Pinkus [58].

Definition 34 (Superpositions of Generalized Ridge Functions)**.** Given a linear map $\phi : \mathbb{R}^n \to \mathbb{R}^d$, a *generalized ridge functions* is an element in

$$\mathcal{M}(\phi) := \left\{ f \circ \phi \mid f \in \mathcal{C}(\mathbb{R}^d) \right\} \subseteq \mathcal{C}(\mathbb{R}^n).$$

Given $\Omega \subseteq \mathbb{R}^{d \times n}$, a *superposition of generalized ridge functions* is an element in

$$\mathcal{M}(\Omega) := \operatorname{Span} \left\{ f \circ \phi \mid f \in \mathcal{C}(\mathbb{R}^d),\, \phi \in \Omega \right\}.$$

If Ω is finite, say $\Omega = \{\phi_i\}_{i \in I}$, we may write

$$\mathcal{M}(\Omega) = \mathcal{M}(\phi_i)_{i \in I} := \left\{ x \mapsto \sum_{i \in I} f_i \circ \phi_i(x) \mid f_i \in \mathcal{C}(\mathbb{R}^d) \right\},$$

or simply write $\mathcal{M}(\phi_1, \ldots, \phi_l)$ when $\Omega = \{\phi_1, \ldots, \phi_l\}$.

To facilitate our exposition, we introduce the following auxiliary notation. Let $A \in \mathbb{R}^{d \times n}$ be matrix, and write it as

$$A := \begin{bmatrix} a_1 \\ \vdots \\ a_d \end{bmatrix},$$

where a_is are the rows of A. Define

$$L(A) := \operatorname{Span} \{a_1, \ldots, a_d\}.$$

Let $\Omega \subseteq \mathbb{R}^{d \times n}$ be a finite set of matrices. Define

$$L(\Omega) := \bigcup_{A \in \Omega} L(A).$$

In the following, we will use the following fundamental result (see [58], p. 65).

Theorem 35. *Let $\Omega = \{A_1, \dots, A_s\} \subseteq \mathbb{R}^{d\times n}$ be a finite set of matrices. Then*

$$\overline{\mathcal{M}(\Omega)} = \overline{\mathcal{M}(L(\Omega))} = \overline{\mathcal{M}(L(A_1) \cup \cdots \cup L(A_s))}.$$

We can characterize the previous sets using the following notions.

Definition 36. Given $\Omega \subseteq \mathbb{R}^n$, define the ideal of polynomials vanishing on Ω as

$$\mathcal{I}(\Omega) := \{p \in \mathbb{R}[x_1, \dots, x_n] \mid p(x) = 0 \, \forall x \in \Omega\},$$

and then, define

$$\mathcal{C}(\Omega) := \{p \in \mathbb{R}[x_1, \dots, x_n] \mid q(D)p = 0 \, \forall q \in \mathcal{I}(\Omega)\}.$$

Theorem 37 (Theorem 6.9 of [58])**.** *In the topology of uniform convergence on compact subsets*

$$\overline{\mathcal{M}(\Omega)} = \overline{\mathcal{C}(\Omega)}.$$

We can compare the closure of spaces of superposition thanks to the following theorem.

Theorem 38. *Let Ω and Ω' be two subsets of $\mathbb{R}^n$ closed under scalar multiplication. If $\mathcal{C}(\Omega) \subsetneq \mathcal{C}(\Omega')$, then $\overline{\mathcal{C}(\Omega)} \subsetneq \overline{\mathcal{C}(\Omega')}$ in topology of uniform convergence on compact sets.*

Proof. If $\mathcal{C}(\Omega) \subsetneq \mathcal{C}(\Omega')$ then there exist $p' \in \mathcal{C}(\Omega')$ and $q \in \mathcal{I}(\Omega)$ such that

$$q'(D) \cdot p' = 0, \ \forall q' \in \mathcal{I}(\Omega')$$

and

$$q(D) \cdot p' \neq 0$$

for each $p \in \mathcal{C}(\Omega)$. Note that $q(D)$ is a continuous operator in the space of tempered distributions and $C(\Omega) \subseteq \ker q(D)$. Since $\ker q(D)$ is a closed subspace by Lemma 39, then $\overline{C(\Omega)} \subseteq \ker q(D)$ while $p' \notin \ker q(D)$, concluding the proof. □

Lemma 39. *Let $(p_n)_{n\in\mathbb{N}}$ be a sequence of polynomials in d variables, each of arbitrary degree, that converges uniformly on compact subsets to a polynomial p. Let $P(\partial_1, \dots, \partial_d)$ be a linear differential operator with constant coefficients, that is, P is a polynomial in d variables. If*

$$P(\partial_1, \dots, \partial_d)\, p_n = 0 \quad \textit{for all } n \in \mathbb{N},$$

then

$$P(\partial_1, \dots, \partial_d)\, p = 0.$$

Proof. Define:

$$\langle f, g\rangle := \int_{\mathbb{R}^n} f(x)g(x)dx.$$

Let ϕ be a smooth function with support on a compact K. We know:

$$\langle p_n, \phi\rangle \to \langle p, \phi\rangle,$$

for $n \to \infty$. Let $Q(\partial_1, \dots, \partial_d)$ be the adjoint operator of $P(\partial_1, \dots, \partial_d)$. This operator is still a linear differential operator when defined on smooth functions with compact support. In particular, $Q(\partial_1, \dots, \partial_d)\phi$ is still a smooth function with support on K. Moreover,

$$\langle p_n, Q(\partial_1, \dots, \partial_d)\phi\rangle = -\langle P(\partial_1, \dots, \partial_d)p_n, \phi\rangle = 0 \tag{8}$$

for each n. Due to convergence on compacts and knowing that the support of $Q(\partial_1, \dots, \partial_d)\phi$ is K, we obtain

$$\langle p_n, Q(\partial_1, \dots, \partial_d)\phi\rangle \to \langle p, Q(\partial_1, \dots, \partial_d)\phi\rangle, \tag{9}$$

for $n \to \infty$. Thanks to Eq. 8 e 9 we get:

$$\langle p, Q(\partial_1, \dots, \partial_d)\phi\rangle = 0.$$

Finally,

$$\langle P(\partial_1, \dots, \partial_d)p, \phi\rangle = -\langle p, Q(\partial_1, \dots, \partial_d)\phi\rangle = 0.$$

Since ϕ is an arbitrary smooth function with compact support, we get

$$\langle P(\partial_1, \dots, \partial_d)p, \phi\rangle = 0$$

for each ϕ with compact support. For the fundamental theorem of calculus of variations, $P(\partial_1, \dots, \partial_d)p$ is identically zero. □

D Proofs and Auxiliary Results

In this section we will concentrate on a particular subset of superpositions of ridge functions, namely, the symmetric ones.

Definition 40 (Symmetric Superpositions)**.** Let $\phi_1, \ldots, \phi_\ell : \mathbb{R}^n \to \mathbb{R}^d$ be linear maps. We define symmetric superpositions of ridge functions as follows:

$$\Delta(\phi_1, \ldots, \phi_\ell) := \left\{ x \mapsto f \circ \phi_1(x) + \cdots + f \circ \phi_\ell \mid f \in \mathcal{C}(\mathbb{R}^d) \right\}.$$

Proposition 41. *The family of functions approximated by $\mathcal{U}_\sigma(M, N)$ coincides with the class $\Delta(\phi_1, \ldots, \phi_\ell)$, where $\phi_1, \ldots, \phi_\ell$ are the basis maps associate to M.*

Proof of Proposition 41. In the general setting, write the linear parts of M and N respectively as $\lambda(M) = \operatorname{Span}\left\{\phi^1, \ldots, \phi^m\right\}$ and $\lambda(N) = \operatorname{Span}\left\{x \mapsto \mathbb{1}^t \cdot x\right\}$. Elements in M_h can be represented as affine maps $x \mapsto Bx + c$ where B and c have the following block representations

$$B = \begin{bmatrix} b_{1,1}\phi^1 + \cdots + b_{1,m}\phi^m \\ \vdots \\ b_{h,1}\phi^1 + \cdots + b_{h,m}\phi^m \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} c_1 \, \mathbb{1} \\ \vdots \\ c_h \, \mathbb{1} \end{bmatrix}.$$

While elements in ${}_hN$ can be represented as affine maps $x \mapsto Ax + d$ where $d \in \mathbb{R}$ and

$$A = \begin{bmatrix} a_1 \, \mathbb{1}^t \\ \vdots \\ a_h \, \mathbb{1}^t \end{bmatrix}.$$

Denote by ϕ_i^j the projection of the i-th component of the function ϕ^j. We can write elements $\eta \in \mathcal{N}_\sigma(M_{h,h}\, N)$ as

$$\eta(x) = A\sigma(Bx + c) = \sum_{j=1}^{h} a_j \sum_{i \in Y} \sigma\left(\sum_{t=1}^{m} b_{j,t}\phi_i^t(x) + c_j \right)$$

for some $a_i, b_{j,t}, c_j \in \mathbb{R}$. But note that

$$\eta(x) = \sum_{j=1}^{h} a_j \sum_{i \in Y} \sigma\left(\sum_{t=1}^{m} b_{j,t}\phi_i^t(x) + c_j \right) = \tag{10}$$

$$\sum_{i \in Y} \sum_{j=1}^{h} a_j \sigma\left(\sum_{t=1}^{m} b_{j,t}\phi_i^t(x) + c_j \right) = \sum_{i \in Y} \zeta(\phi_i^1(x), \ldots, \phi_i^m(x)) \tag{11}$$

where

$$\zeta(y_1, \ldots, y_l) := \sum_{j=1}^{h} a_j \sigma\left(\sum_{t=1}^{m} b_{j,t} y_t + c_j \right)$$

is a standard multilayer perceptron in $\mathcal{N}_\sigma(\mathbb{R}^l, \mathbb{R}^h, \mathbb{R})$. Since, the the multilayer perceptron is universal, thanks to (10) we can approximate any superposition in $\Delta(\phi_1, \ldots, \phi_l)$. Thus, we have

$$\overline{\Delta(\phi_1, \ldots, \phi_l)} \subseteq \mathcal{U}_\sigma(M, N).$$

On the other hand, by (10),

$$\bigcup_{h \in \mathbb{N}} \mathcal{N}_\sigma(M_{h,h}\, N) \subseteq \Delta(\phi_1, \ldots, \phi_l).$$

It follows that their closures coincide, which concludes the proof. □

Let M be a vector space of affine maps such that $\lambda(M) = \operatorname{Span}\left\{\phi^1, \ldots, \phi^m\right\}$, and let N be the set of invariant affine maps. Denote $\rho = \rho(\mathcal{N}_\sigma(M, N))$. We denote by $\{\{x_1, \ldots, x_n\}\}$ the multiset of elements $x_1, \ldots, x_n$. We have the following proposition.

Proposition 42. *With the notation defined above, we have* $(x, y) \in \rho(\mathcal{N}_\sigma(M, N)) = \rho(\mathcal{U}_\sigma(M, N))$ *if and only if*

$$\{\{\phi_1(x), \dots, \phi_\ell(x)\}\} = \{\{\phi_1(y), \dots, \phi_\ell(y)\}\},$$

where we identify Y *with* $[\ell]$*, which inherits its* G*-set structure from* Y*, and the maps* ϕ_i *are those defined in* (6)*.*

Proof. By the combination of Proposition 41 and Theorem 8 in [18], we have $\rho(\mathcal{N}_\sigma(M, N)) = \rho(\mathcal{U}_\sigma(M, N) = \rho(\Delta(\phi_1, \dots, \phi_\ell))$. Thus, it suffices to verify this property for $\Delta(\phi_1, \dots, \phi_\ell)$. Note that if x and y satisfy

$$\{\{\phi_1(x), \dots, \phi_\ell(x)\}\} = \{\{\phi_1(y), \dots, \phi_\ell(y)\}\},$$

then, for each $F \in \Delta(\phi_1, \dots, \phi_\ell)$, we have

$$F(x) = f \circ \phi_1(x) + \cdots + f \circ \phi_\ell(x) = f \circ \phi_1(y) + \cdots + f \circ \phi_\ell(y) = F(y).$$

On the other hand, if

$$\{\{\phi_1(x), \dots, \phi_\ell(x)\}\} \neq \{\{\phi_1(y), \dots, \phi_\ell(y)\}\},$$

then we have two possibilities: either there exists an i such that $\phi_i(x) \neq \phi_i(y)$, or there exists a value γ such that the number of indices i with $\phi_i(x) = \gamma$ (denoted s) differs from the number of indices i with $\phi_i(y) = \gamma$ (denoted t). In the first case, we can choose an interpolating function f that does not vanish at $\phi_i(x)$ and is zero on the other values in consideration. In this case,

$$F(x) = f \circ \phi_1(x) + \cdots + f \circ \phi_\ell(x) \neq 0 = f \circ \phi_1(y) + \cdots + f \circ \phi_\ell(y) = F(y).$$

In the other case, we can similarly chose a function f nonzero on γ and zero on all the other values in consideration. In this case,

$$F(x) = f \circ \phi_1(x) + \cdots + f \circ \phi_\ell(x) = sf(\gamma) \neq tf(\gamma) = f \circ \phi_1(y) + \cdots + f \circ \phi_\ell(y) = F(y).$$

This concludes the proof. □

Proposition 9 follows directly from Proposition 42.

Proof of Proposition 9. Note that, by Theorem 6, $\rho(\mathcal{U}_\sigma(C^1, I)) = \mathcal{C}_{S_n}(\mathbb{R}^n)$ and thus has maximal separation power in the context of permutation invariance; that is, it separates two points if and only if they lie in the same S_n-orbit. Note that the basis maps associated to C^1 are $e_1^\top, \dots, e_n^\top$. Hence, by Proposition 42, $(x, y) \in \rho(\mathcal{U}_\sigma(C^1, I))$ if and only if $\{\{x_1, \dots, x_n\}\} = \{\{y_1, \dots, y_n\}\}$. This holds if and only if x and y lie in the same S_n-orbit. Thus, $\mathcal{U}_\sigma(C^1, I)$ also has maximal separation power, and hence

$$\rho(\mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^G, \mathbb{R})) = \rho(\mathcal{U}_\sigma(C^1, I)).$$

Since

$$\mathcal{U}_\sigma(C^1, I) \subseteq \mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^n, \mathbb{R}) \subseteq \mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^G, \mathbb{R}),$$

it follows that

$$\rho(\mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^G, \mathbb{R})) \subseteq \rho(\mathcal{U}_\sigma(\mathbb{R}^n, \mathbb{R}^n, \mathbb{R})) \subseteq \rho(\mathcal{U}_\sigma(C^1, I)).$$

Therefore, all inclusions must be equalities. □

Proof of Theorem 13. We start by proving that $\mathcal{M}(\phi_1, \dots, \phi_l)^G = \Delta(\phi_1, \dots, \phi_l)$. Indeed, consider $\mathcal{R} : \mathcal{C}(\mathbb{R}^n) \to \mathcal{C}(\mathbb{R}^n)^G$ the Reynolds operator. For each $F \in \mathcal{M}(\phi_1, \dots, \phi_l)$,

$$\mathcal{R}(F)(x) = \sum_{g \in G} F(gx) = \sum_{g \in G} f_1 \circ \phi_1(gx) + \cdots + f_l \circ \phi_l(gx) =$$
$$M \cdot [(f_1 \circ \phi_1(x) + \cdots + f_1 \circ \phi_l(x)) + \cdots + (f_l \circ \phi_1(x) + \cdots + f_l \circ \phi_l(x))] =$$
$$M \cdot [(f_1 + \cdots + f_l) \circ \phi_1(x) + \cdots + (f_1 + \cdots + f_l) \circ \phi_l(x)] \in \Delta(\phi_1, \dots, \phi_l),$$

where we denote $M := |\operatorname{Stab}(\phi_1)| = \cdots = |\operatorname{Stab}(\phi_l)|$.
Furthermore, the map

$$\mathcal{C}(\mathbb{R}^d) \times \cdots \times \mathcal{C}(\mathbb{R}^d) \to \mathcal{C}(\mathbb{R}^d)$$
$$(f_1, \dots, f_l) \mapsto f_1 + \cdots + f_l$$

is surjective; hence the Reynolds operator is surjective as a map from $\mathcal{M}(\phi_1, \ldots, \phi_l)$ to $\Delta(\phi_1, \ldots, \phi_l)$. This proves the desired equality.

If f is an invariant function, we have by Theorem 37 and Theorem 41:

$$\begin{aligned} & f \in \mathcal{U}_\sigma(M, N) \iff f \in \overline{\Delta(\phi_1, \ldots, \phi_\ell)} \\ \iff & f \in \overline{\mathcal{M}(\phi_1, \ldots, \phi_\ell)} \iff f \in \overline{\mathcal{C}(\phi_1, \ldots, \phi_\ell)} \\ \iff & P(\partial_1, \ldots, \partial_n) f = 0 \quad \text{for each } P \in \mathcal{I}(L(\phi_1) \cup \cdots \cup L(\phi_\ell)). \end{aligned}$$

This concludes the proof. □

Proof of Theorem 14. The final part of the proof of Theorem 13 implies that if $f \in \mathcal{U}_\sigma(M, N)$, then for any $P \in \mathcal{I}(L(\phi_1) \cup \cdots \cup L(\phi_\ell))$, $P(\partial_1, \ldots, \partial_n) f = 0$. By Remark 33, we know

$$\mathcal{I}(L(\phi_1) \cup \cdots \cup L(\phi_\ell)) = \mathcal{I}(L(\phi_1)) \cap \cdots \cap \mathcal{I}(L(\phi_\ell)) \supseteq \mathcal{I}(L(\phi_1)) \cdots \mathcal{I}(L(\phi_\ell)).$$

For any $\alpha = 1, \ldots, \ell$ and arbitrary $c_\alpha \in \ker \phi_\alpha^\top$, note that for

$$c_\alpha^\top x \in \mathcal{I}(L(\phi_\alpha)).$$

Hence,

$$(c_1^\top x) \cdots (c_\ell^\top x) \in \mathcal{I}(L(\phi_1)) \cdots \mathcal{I}(L(\phi_\ell)).$$

Whose associated differential operator can be written as $D_{c_1} \cdots D_{c_\ell}$. Therefore,

$$D_{c_1} \cdots D_{c_\ell} f = 0,$$

concluding the proof. □

Proof of Theorem 15. By Proposition 42, separation-constrained universality is equivalent to the ability to approximate any function of the form $F(\phi_1, \ldots, \phi_\ell)$, where F is continuous and S_ℓ-invariant.

Recall that the basis maps are defined as

$$\phi_i = (\phi_i^1, \ldots, \phi_i^m).$$

Let $W = \mathbb{R}^Y$ for some finite G-set Y. Since $M = \mathrm{Aff}_G(V, \mathbb{R}^Y)$, we can, for a suitable choice of basis, select elements $\alpha_i \in Y$ such that $\phi_i^1 = e_{\alpha_i}^\top$ for each $i = 1, \ldots, \ell$.

In particular, the function

$$F : x \mapsto G(e_{\alpha_1}^\top x, \ldots, e_{\alpha_\ell}^\top x),$$

for some $G : \mathbb{R}^\ell \to \mathbb{R}$, is one that should be approximable under separation constraints.

Specifically, we define G as the symmetrization of the monomial

$$M(x_1, \ldots, x_\ell) = x_1^{a_1} \cdots x_\ell^{a_\ell},$$

that is,

$$G(x_1, \ldots, x_\ell) = \sum_{\sigma \in S_\ell} M(x_{\sigma(1)}, \ldots, x_{\sigma(\ell)}).$$

Now, observe that if

$$D_{c_1} \cdots D_{c_\ell} M \neq 0,$$

then

$$D_{c_1} \cdots D_{c_\ell} G \neq 0$$

for any choice of $c_i \in \ker \phi_i$ for some $i = 1, \ldots, \ell$. Therefore, F cannot be approximated by $\bigcup_{h \in \mathbb{N}} \mathcal{N}(M_h, h\, N)$.

This follows because the differential operator $D_{c_1} \cdots D_{c_\ell}$ reduces the degree of each monomial in G by at most ℓ. Thanks to the hypothesis $a_i + \ell < a_{i+1}$ for each $i = 1, \ldots, \ell$, and $a_1 > \ell$, all resulting monomials in $D_{c_1} \cdots D_{c_\ell} G$ have distinct multidegrees. In particular, $D_{c_1} \cdots D_{c_\ell} M$, being one of these monomials and being nonzero, implies that $D_{c_1} \cdots D_{c_\ell} G$ is itself nontrivial.

This proves that if $D_{c_1} \cdots D_{c_\ell} M \neq 0$, then the function F cannot be approximated by $\bigcup_{h \in \mathbb{N}} \mathcal{N}(M_{h,h}\, N)$.

By direct computation, the coefficients of the monomials of multidegree $(a_1 - s_1, \ldots, a_\ell - s_\ell)$ in $D_{c_1} \cdots D_{c_\ell} M$ are given by

$$\sum_{\sigma \in S_\ell} \frac{a_{i_1}!}{s_{i_1}!} \cdots \frac{a_{i_r}!}{s_{i_r}!} (c_{\sigma(1),1} \cdots c_{\sigma(s_1),1}) \cdot (c_{\sigma(s_1+1),2} \cdots c_{\sigma(s_1+s_2),2}) \cdots (c_{\sigma(\ell - s_\ell),\ell} \cdots c_{\ell,\ell}).$$

where $s_1, \ldots, s_\ell \in \{0, \ldots, \ell\}$, $s_1 + \cdots + s_\ell = \ell$ and $i_1, \ldots, i_r$ are the indices such that $s_{i_j} \neq 0$.

If at least one of these coefficients is nonzero, then $D_{c_1} \cdots D_{c_\ell} F$ is nontrivial and thus cannot be approximated by $\bigcup_{h \in \mathbb{N}} \mathcal{N}(M_{h,h}\, N)$. □

Proof of Theorem 18. Define V, W, and $\iota : V \to W$ as in Corollary 44, which states that

$$\mathcal{N}(V, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z) = \iota^* \mathcal{N}(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z)$$

for each $h \in \mathbb{N}$.

Since H is normal in G, the quotient G/H is a group and the action of H on W is trivial, W is a G/H-representation, and we have the identification $\mathcal{C}_G(W, Z) = \mathcal{C}_{G/H}(W, Z)$.

From Ravanbakhsh [5], it is known that shallow equivariant neural networks with the regular representation as input are universal approximators. In this case,

$$\bigcup_{h \in \mathbb{N}} \mathcal{N}(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z)$$

is universal in $\mathcal{C}_G(W, Z) = \mathcal{C}_{G/H}(W, Z)$.

Furthermore, the pullback map $\iota^* : \mathcal{C}(V, Z) \to \mathcal{C}(W, Z)$ is a continuous linear operator. Hence,

$$\begin{aligned}
\overline{\bigcup_{h \in \mathbb{N}} \mathcal{N}(V, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z)} &= \overline{\bigcup_{h \in \mathbb{N}} \iota^* \mathcal{N}(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z)} \\
&= \overline{\iota^* \left(\bigcup_{h \in \mathbb{N}} \mathcal{N}(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z) \right)} \\
&= \iota^* \left(\overline{\bigcup_{h \in \mathbb{N}} \mathcal{N}(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z)} \right) \\
&= \iota^* \left(\mathcal{C}_{G/H}(W, Z) \right) = \iota^* \left(\mathcal{C}_G(W, Z) \right).
\end{aligned}$$

Therefore, the left-hand side is equivariant-universal as well. Finally, observe that $\iota^*(\mathcal{C}_G(W, Z))$ is an algebra of functions containing the constants, so it is separation-constrained universal by the Stone–Weierstrass theorem. □

Lemma 43. *Let H be normal subgroup of G and K an arbitrary subgroup of G. Consider the standard immersion map*

$$\iota : \mathbb{R}^{G/HK} \to \mathbb{R}^{G/K}$$

as the standard injection induced by the subgroup inclusion $K < KH$. We define the pullback map

$$\begin{aligned}
\iota^* : \mathcal{C}(\mathbb{R}^{G/K}, Z) &\to \mathcal{C}(\mathbb{R}^{G/HK}, Z) \\
f &\mapsto f \circ \iota
\end{aligned}$$

for any G-representation Z.

Proof. Note that $\iota^* \operatorname{Hom}_G(\mathbb{R}^{G/K}, \mathbb{R}^{G/H}) \subseteq \operatorname{Hom}_G(\mathbb{R}^{G/HK}, \mathbb{R}^{G/H})$, since ι^* is linear and preserves equivariance. Moreover, since ι is injective, the induced map ι^* is surjective.

Now, assume that H is normal. Then,

$$\dim \operatorname{Hom}_G(\mathbb{R}^{G/K}, \mathbb{R}^{G/H}) = |H \backslash G / K| = |H \backslash G / HK| = \dim \operatorname{Hom}_G(\mathbb{R}^{G/HK}, \mathbb{R}^{G/H}).$$

This equality of dimensions, together with the inclusion and surjectivity above, implies that ι^* is an isomorphism of vector spaces. In particular,

$$\iota^* \operatorname{Hom}_G(\mathbb{R}^{G/K}, \mathbb{R}^{G/H}) = \operatorname{Hom}_G(\mathbb{R}^{G/HK}, \mathbb{R}^{G/H}).$$

□

Corollary 44. *Let $V = \mathbb{R}^{G/K_1} \oplus \cdots \oplus \mathbb{R}^{G/K_d}$ and define $W = \mathbb{R}^{G/K_1 H} \oplus \cdots \oplus \mathbb{R}^{G/K_d H}$. Consider the standard immersion map $\iota : W \to V$ as the standard injection defined component by component and induced by the subgroup inclusion $K_i < K_i H$ for $i = 1, \ldots, d$. We define the pullback map*

$$\begin{aligned} \iota^* : \mathcal{C}(V, Z) &\to \mathcal{C}(W, Z) \\ f &\mapsto f \circ \iota \end{aligned}$$

for any G-representation Z. Then

$$\mathcal{N}(V, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z) = \iota^* \mathcal{N}(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z),$$

for any G-representation Z.

Proof. By the properties of representation homomorphisms under direct sums, we have

$$\begin{aligned} \operatorname{Hom}_G(V, \mathbb{R}^{G/H} \otimes \mathbb{R}^h) &= \operatorname{Hom}_G\left(\mathbb{R}^{G/K_1} \oplus \cdots \oplus \mathbb{R}^{G/K_d}, \mathbb{R}^{G/H} \otimes \mathbb{R}^h\right) \\ &= \bigoplus_{i=1}^{d} \operatorname{Hom}_G(\mathbb{R}^{G/K_i}, \mathbb{R}^{G/H})^{\oplus h}. \end{aligned}$$

By the definition of ι and Lemma 43, it follows that

$$\iota^* \operatorname{Hom}_G(V, \mathbb{R}^{G/H} \otimes \mathbb{R}^h) = \operatorname{Hom}_G(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h)$$

for each $h \in \mathbb{N}$. Consequently,

$$\iota^* \operatorname{Aff}_G(V, \mathbb{R}^{G/H} \otimes \mathbb{R}^h) = \operatorname{Aff}_G(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h)$$

for every $h \in \mathbb{N}$ as well.

Therefore, for any G-representation Z, we obtain

$$\mathcal{N}(V, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z) = \iota^* \mathcal{N}(W, \mathbb{R}^{G/H} \otimes \mathbb{R}^h, Z),$$

since ι is precomposed with the input in the first layer. □