ON UNIVERSALITY OF DEEP EQUIVARIANT NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Universality results for equivariant neural networks remain rare. Those that do exist typically hold only in restrictive settings: either they rely on regular or higher-order tensor representations, leading to impractically high-dimensional hidden spaces, or they target specialized architectures, often confined to the invariant setting. This work develops a more general account. For invariant networks, we establish a universality theorem under separation constraints, showing that the addition of a fully connected readout layer secures approximation within the class of separation-constrained continuous functions. For equivariant networks, where results are even scarcer, we demonstrate that standard separability notions are inadequate and introduce the sharper criterion of *entry-wise separability*. We show that with sufficient depth or with the addition of appropriate readout layers, equivariant networks attain universality within the entry-wise separable regime. Together with prior results showing the failure of universality for shallow models, our findings identify depth and readout layers as a decisive mechanism for universality, additionally offering a unified perspective that subsumes and extends earlier specialized results.

1 INTRODUCTION

Symmetry has emerged as a key organizing principle in deep learning. Equivariant neural networks encode symmetry by ensuring that transformations of the input are mirrored by corresponding transformations of the output. This inductive bias has proven successful in applications ranging from vision and molecular modeling to representation learning on graphs and manifolds (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bronstein et al., 2021).

An important concern, however, is whether the introduced inductive biases may impose additional undesired constraints beyond symmetry. In this direction, the majority of work focuses on the study of *expressivity*, a broad notion that intuitively reflects the capacity of a family of models to represent or approximate arbitrarily complex target functions. This notion admits different formalizations, but two main approaches are currently investigated in the literature. The first directly tackles *universality*, understood here as the problem of approximating all symmetry-compatible target functions (Ravanbakhsh, 2020; Keriven & Peyré, 2019; Maron et al., 2019b; Sonoda et al., 2022), which, however, often requires models with impractically large intermediate representations. The second approach concerns the ability of models to distinguish input pairs, their *separation power*, an ability that also constrains the functions they can approximate. Separation has been extensively studied in the graph learning community through the lens of the Weisfeiler–Leman test (Morris et al., 2019; Maron et al., 2019a), and more recently in a general equivariant setting (Joshi et al., 2023; Pacini et al., 2024b). In particular, Chen et al. (2019) and Joshi et al. (2023) present initial approaches to universality that explicitly account for separation constraints. They establish universality results up to Weisfeiler–Leman and orbit separation, respectively, thereby furnishing the first cases of *separation-constrained universality*.

However, Pacini et al. (2025) suggest a more nuanced landscape for the interaction between separation and universality. For instance, they present examples of invariant shallow architectures with the same separation power but different approximation power, showing that although separation is a *necessary* condition for approximation, it may fail to be a *sufficient* one. Nevertheless, Zaheer et al. (2017), Qi et al. (2017), and Segol & Lipman (2020) show that adding fully connected readout layers or increasing the depth of this limited class of architectures transforms them into universal models up to separation. This suggests that depth and readout layers may play a crucial role in achieving separation-constrained universality and, more generally, in efficiently enhancing

approximation power. In this paper, we shed light on this phenomenon by investigating the role of depth in separation-constrained universality, both in the invariant and equivariant regimes, and offer a unified framework that goes beyond earlier architecture-specific results. Our first result is a *separation-constrained universality theorem* for invariant networks, showing that models with fully connected readouts can approximate every continuous function consistent with their separation relation (Section 4). We then turn to the equivariant setting, where a simple example shows that standard separability is too coarse to characterize universality. To address this, we introduce the notion of *entry-wise separability* (Section 5.1). Intuitively, instead of considering the separation relation of the entire function, we examine all the separation relations of its projections onto individual output coordinates simultaneously. With this notion in place, we prove two *entry-wise separation-constrained universality theorems*. These results establish that deep equivariant networks achieve universality either when depth is sufficient to stabilize separation or when specific output layers act, in the equivariant setting, as surrogates of fully connected readouts in the invariant case (Section 5.2). In summary, our results identify depth and readouts as key factors for universality across broad classes of invariant and equivariant architectures. They clarify the role of separation in approximation and subsume earlier results restricted to shallow or architecture-specific settings.

We summarize our main contributions below:

- We establish a *separation-constrained universality theorem* for invariant networks (Theorem 1), showing that the addition of a fully connected readout guarantees approximation within the separation-constrained class.
- We introduce the concept of *entry-wise separability* and demonstrate, via Example 3, that standard separability fails to capture the universality class of equivariant networks.
- Building on this refinement, we prove two *entry-wise separation-constrained universality theorems*, showing that equivariant networks achieve universality either once entry-wise separation relations stabilize with depth (Theorem 2), or when equipped with specific readout layers (Theorem 3).

2 Related Work

Equivariant architectures have emerged as a principled framework to incorporate symmetry into machine learning (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bronstein et al., 2021). Beyond convolution, a variety of techniques have been developed to enforce equivariance in over-parameterized or hierarchical representation-learning mechanisms, including approximate equivariance (Finzi et al., 2021; Petrache & Trivedi, 2023), tensor- and polynomial methods (Thomas et al., 2018), and hybrid polynomial models (Dym & Gortler, 2022). These approaches have been extended to different data structures, such as point clouds (Fuchs et al., 2020), graphs (Victor Garcia Satorras et al., 2021), and simplicial complexes (Battiloro et al., 2025). Thanks to this versatility, these models were able to adapt to diverse symmetry-sensitive domains, including high-energy physics (Bogatskiy et al., 2020), structural biology and drug discovery (Jumper et al., 2021), robotics (Huang et al., 2023), and medical imaging (Lafarge et al., 2021).

However, due to this heterogeneity of the landscape, a principled understanding of how such biases affect models remains fragmented and far from complete. Most work focuses on *expressivity*, the capacity of a model class to represent arbitrarily complex target functions. Ravanbakhsh (2020) and Sonoda et al. (2022) show that shallow architectures with regular hidden representations can approximate any equivariant map. However, these results require hidden spaces whose dimension scales with group size, making them impractical. Similarly, Maron et al. (2019b) establish universality for invariant networks with high-order tensor representations, though such constructions remain far from practical use. Among architectures used in practice, graph neural networks occupy a prominent role in the geometric deep learning literature. The study of their expressivity has been carried out primarily through the lenses of separation, via the Weisfeiler-Leman test, which enables fine-grained understanding of the model's separation capabilities. This focus is justified by the result of Chen et al. (2019), who establish universality up to Weisfeiler–Leman separation. Building on this, Joshi et al. (2023) extended the result to geometric graph neural networks, establishing universality up to orbit separation. While these universality results remained confined to the invariant setting, the separation power of equivariant networks is now well characterized (Geerts, 2020; Geerts & Reutter, 2022; Pacini et al., 2024b), and depth plays a central role. However, Pacini et al. (2025) recently showed that

depth and additional readout layers play a more subtle role in approximation relative to separation, demonstrating that they can change the class of functions that are approximable without altering separation. This stands in stark contrast to the theory of classical neural networks, where depth is known to improve parameter efficiency but not the class of approximable functions (Telgarsky, 2016; Yarotsky, 2017; 2018).

We extend this literature in two ways. First, we prove that adding a fully connected readout layer is sufficient to achieve separation-constrained universality for invariant neural networks. Second, we introduce and analyze *entry-wise separability*, showing that it provides the appropriate refinement for extending separation-constrained universality to equivariant networks. We present these results in a mathematical framework that is general enough to encompass prior settings while precisely capturing the underlying phenomena.

3 Preliminaries

3.1 Groups and Equivariance

We study functions that behave consistently under prescribed transformations. Among them, some are naturally formalized by the algebraic structure of a *group*: sets of transformations closed under composition and containing inverses and an identity. While group theory provides the natural framework for reasoning about symmetry, in the context of neural networks it is convenient to reformulate these ideas in linear-algebraic terms. This is achieved through *representation theory*, which encodes abstract group elements as matrix actions on vector spaces (Serre, 1977).

Permutation representations play a central role in this work. They arise when a group G acts on a finite set X, where the action is given by an identification of G as a subset of permutations of X. Let $\mathbb{R}^X$ denote the space of real-valued functions on X, and for each $x \in X$ define the indicator $e_x \in \mathbb{R}^X$ by $e_x(x) = 1$ and $e_x(y) = 0$ for $y \neq x$. The collection $\{e_x\}_{x\in X}$ forms a canonical basis of $\mathbb{R}^X$. The associated permutation representation is the linear action on $V = \mathbb{R}^X$ given by $g(e_x) = e_{gx}$, where gx is the result of g acting on x for $g \in G, x \in X$.

If V and W are permutation representations of G, a map $\phi : V \to W$ is called *G-equivariant* when $\phi(gv) = g\phi(v)$ for all $g \in G$, $v \in V$. We denote by $\mathrm{Hom}(V, W)$ the space of linear maps and by $\mathrm{Hom}_G(V, W)$ the subspace of G-equivariant linear maps. Similarly, $\mathrm{Aff}(V, W)$ denotes the space of affine maps and $\mathrm{Aff}_G(V, W) \subseteq \mathrm{Aff}(V, W)$ the subset of G-equivariant affine maps. All these spaces are real vector spaces under pointwise addition and scalar multiplication.

Further preliminaries are provided in Appendix A.

3.2 Layer Spaces, Neural Spaces & Equivariant Neural Networks

We now describe equivariant neural networks, which are model classes with group equivariant layers. Throughout, we restrict attention to networks equivariant under the action of a finite group, with layers given by permutation representations and equipped with arbitrary point-wise continuous activations. We begin by introducing the notion of a *layer space*, namely a space of affine maps subject to additional constraints—such as equivariance requirements or restrictions on the set of permissible filters—which will serve as the fundamental building block of the neural architectures under consideration.

Definition 1 (Layer Spaces)**.** *Let G be a finite group acting on a finite set X, let $\mathbb{R}^X$ be the permutation representation associated with this action, and let V be another permutation representation of G. A* layer space *is a subset $M \subseteq \mathrm{Aff}_G(V, \mathbb{R}^X)$. In this work, we focus on spaces of 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 $\phi^1, \ldots, \phi^k \in \mathrm{Hom}_G(V, \mathbb{R}^X)$, the sets $X_1, \ldots, X_\ell$ denote all the orbits of X under the G-action, and $\mathbb{1}_{X_j} := \sum_{x\in X_j} e_x$ for $j = 1, \ldots, \ell$, with $\{e_x\}_{x\in X}$ the canonical basis of $\mathbb{R}^X$.

Example 1. *We give some examples of layer spaces, L, I, C, and P, which will be used throughout the manuscript as running references in our analysis of the universality phenomena. These layer*

spaces correspond to widely used architectures in geometric deep learning and illustrate how standard models naturally fit into the general form (1).

(i) ***Linear Layer:*** *Linear layers in standard neural networks are given by elements of* $\mathrm{Aff}(\mathbb{R}^n, \mathbb{R}^m)$. *For the action of any group, we can define the set of affine linear maps between trivial representations* $L := \mathrm{Aff}(\mathbb{R}, \mathbb{R})$, *whose relevance will become clearer later, for instance, in relation to (3).*

(ii) ***Invariant Layer:*** *Let* G *be a finite group acting on a finite set* X, *and let* $\mathbb{R}^X$ *denote the associated permutation representation. We denote by* $\mathbb{R}$ *the trivial real representation of* G. *The space of* G*-invariant affine maps from* $\mathbb{R}^X$ *to* $\mathbb{R}$ *is denoted by* $I := \mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R})$. *If* $X = X_1 \sqcup \cdots \sqcup X_\ell$ *is the orbit decomposition of* X, *then we have the characterization*

$$I := \Big\{ v \mapsto \sum_{i=1}^{\ell} x_i \mathbb{1}_{X_i}^\top \cdot v + y \mid x_1, \ldots, x_\ell, y \in \mathbb{R} \Big\}.$$

(iii) ***Convolutional Layer:*** *Standard convolutional layers correspond to maps equivariant with respect to the cyclic group* $G = \mathbb{Z}_n \times \mathbb{Z}_n$, *acting by the standard cyclic permutations on the product* $X = [n] \times [n]$, *and can be naturally formulated within the framework of permutation representations. Here we consider the generalization to general finite* G *acting on finite* X, *and focus on convolutional layers with filter width* 1 *between general permutation representations* $\mathbb{R}^X$. *These can be written in the form of (1) as follows.*

$$C := \Big\{ v \mapsto x \,\mathrm{id} \cdot v + \sum_{i=1}^{\ell} y_i \mathbb{1}_{X_i} \;\Big|\; x, y_1, \ldots, y_\ell \in \mathbb{R} \Big\}. \tag{2}$$

Note that here $C \subseteq \mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X)$ *for any action of* G *on* X, *where* $X = X_1 \sqcup \cdots \sqcup X_\ell$ *denotes the orbit decomposition of* X. *The same setting can be extended to wider filters, but for ease of exposition, will not be used in this paper.*

(iv) ***PointNet Layer:*** *Sum-pooling PointNet layers (Qi et al., 2017) are designed to process unordered collections, such as point clouds, by enforcing permutation equivariance. In the simplest case, an input configuration of* n *real elements is represented as a vector* $a \in \mathbb{R}^n$, *where we identify* $\mathbb{R}^X = \mathbb{R}^{[n]} \cong \mathbb{R}^n$. *This definition extends analogously to the general case with multi-dimensional features. Equivariant PointNet layers act on such inputs using maps in the space* $\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$. *Zaheer et al. (2017) characterized this space as*

$$P := \Big\{ v \mapsto (x_1 \,\mathrm{id} + x_2 \mathbb{1}\mathbb{1}^\top) \cdot v + y\mathbb{1} \;\Big|\; x_1, x_2, y \in \mathbb{R} \Big\},$$

where $\mathbb{1} = \mathbb{1}_{[n]} = [1, \ldots, 1]^\top$.

We restrict our study to point-wise activations, also referred to in the literature as *component-wise* or *entry-wise* activations.

Definition 2 (Point-wise Activation)**.** *Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be a nonlinear activation. Given a permutation representation* $\mathbb{R}^X$ *of a group* G, *we define the associated* point-wise activation $\tilde{\sigma} : \mathbb{R}^X \to \mathbb{R}^X$ *by* $\tilde{\sigma}\big(\sum_{x \in X} \alpha_x e_x\big) = \sum_{x \in X} \sigma(\alpha_x) e_x$. *Wherever the usage is unambiguous from context, we will denote both* σ *and* $\tilde{\sigma}$ *by the same symbol.*

Throughout the paper, we assume that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is non-polynomial, as we focus on universality results for architectures of fixed depth.

We now state the definition of a neural network and of the functional space of a fixed neural architecture, which we call a *neural space*, also referred to in the literature as a *neuromanifold* (Calin, 2020).

Definition 3 (Neural Networks and Neural Spaces)**.** *Let* G *be a group and* $V_0, \ldots, V_d$ *be permutation representations of* G. *For each* $i = 1, \ldots, d$, *let* M_i *be a layer space in* $\mathrm{Aff}_G(V_{i-1}, V_i)$. *For* $d \geq 2$, *the* neural space *associated with layers* $M_1, \ldots, M_d$ *and activation* σ *is defined recursively by*

$$\mathcal{N}_\sigma(M_1, \ldots, M_d) = \big\{ \phi^d \circ \tilde{\sigma} \circ \cdots \circ \tilde{\sigma} \circ \phi^1 \;\big|\; \phi^i \in M_i \;\text{ for each } i = 1, \ldots, d \big\}.$$

Any $\eta^d \in \mathcal{N}_\sigma(M_1, \ldots, M_d)$ *is called a* neural network *with layers in* $M_1, \ldots, M_d$ *and activation* σ.

3.3 Universality Classes and Separation

We aim to characterize the class of continuous functions approximable by neural networks with fixed architecture. We generalize the notion of *universality class* introduced by Pacini et al. (2025) for shallow networks, to networks of arbitrary depth. Before giving a formal definition, we introduce an auxiliary notion which plays the role of width in classical (non-equivariant) universality results (Pinkus, 1999), since it can be interpreted as the dimension of intermediate invariant feature representations. If V and W are permutation representations of a finite group G and $M \subseteq \mathrm{Aff}_G(V, W)$ is as defined in (1), then, for each $h, k \in \mathbb{N}$ we define $M^{k\times h}$ as the subspace of $\mathrm{Aff}_G(V \otimes \mathbb{R}^k, W \otimes \mathbb{R}^h)$:

$$M^{k\times h} := \left\{ \begin{array}{c} (x_1, \dots, x_k) \mapsto \left(\sum_{j=1}^k f_{1,j}(x_j), \dots, \sum_{j=1}^k f_{h,j}(x_j)\right) \\ f_{ij} \in M, i = 1, \dots, h, j = 1, \dots, k \end{array} \right\}. \tag{3}$$

Example 2. *Recall that the definition of $L = \mathrm{Aff}(\mathbb{R}, \mathbb{R})$, the layer space $L^{k\times h}$ is the set of all affine maps from $\mathbb{R}^k$ to $\mathbb{R}^h$, namely $\mathrm{Aff}(\mathbb{R}^k, \mathbb{R}^h)$. Since $L = \mathrm{Aff}_G(\mathbb{R}, \mathbb{R})$ where G acts trivially on $\mathbb{R}$, this layer space can be interpreted as the space of affine G-equivariant maps between trivial representations. In this sense, the multiplicities k and h correspond to the widths of intermediate representations in the standard neural network setting.*

With the above notation in place, we can now provide a general definition of universality classes. Intuitively, a universality class consists of all functions that can be uniformly approximated on compact sets by neural networks of a given architecture, with variable multiplicities of layer spaces.

Definition 4 (Universality Classes)**.** *The* universality class $\mathcal{U}_\sigma(M_1, \dots, M_d)$ *associated with the layer spaces $M_1, \dots, M_d$ is defined as*

$$\mathcal{U}_\sigma(M_1, \dots, M_d) := \overline{\bigcup_{\vec{h}\in\mathbb{N}^{d-1}} \mathcal{N}_\sigma\big(M_1^{1\times h_1}, M_2^{h_1\times h_2}, M_3^{h_2\times h_3}, \dots, M_{d-1}^{h_{d-2}\times h_{d-1}}, M_d^{h_{d-1}\times 1}\big)},$$

where the overline denotes closure in the topology of uniform convergence on compact sets.

Invariant networks are inherently unable to distinguish between elements in the same group orbits, but additional undesired separability constraints may arise when dealing with neural networks with particular architectures employed in practice. A prominent example is given by graph neural networks, which are known to be subject to separation constraints equivalent to the Weisfeiler–Leman test (Chen et al., 2019). To study the universality classes arising from architectures employed in practice, we must therefore take these separability constraints into account. We will use the following natural definitions of *separation* and of *separation-constrained universality*.

Definition 5 (Separation-Constrained Universality)**.** *Let $\mathcal{U} \subseteq \{f : V \to W\}$ be a family of functions. We say that $\mathcal{U}$ **separates** two points $\alpha, \beta \in V$ if there exists $f \in \mathcal{U}$ such that $f(\alpha) \neq f(\beta)$. The set of pairs that cannot be distinguished by any $f \in \mathcal{U}$ induces an equivalence relation:*

$$\rho(\mathcal{U}) = \big\{(\alpha, \beta) \in V \times V \;\mid\; f(\alpha) = f(\beta) \text{ for all } f \in \mathcal{U}\big\}.$$

*We say that $\mathcal{U}$ is **separation-constrained universal** if it approximates exactly the class of continuous functions that preserve the equivalence relation $\rho = \rho(\mathcal{U})$, namely*

$$\mathcal{C}_\rho(V, W) = \big\{f \in \mathcal{C}(V, W) \;\mid\; f(\alpha) = f(\beta) \text{ whenever } (\alpha, \beta) \in \rho\big\}.$$

Note that separability is a *necessary condition* for uniform approximation on compact sets: any sequence of functions with prescribed separation power ρ converges only to functions that also respect ρ. In other words, $\mathcal{C}_\rho(V, W)$ is a closed subset of $\mathcal{C}(V, W)$ in the topology of uniform convergence on compact sets.

As noted in Sections 1 and 2, the literature on universality for equivariant neural networks is typically architecture-dependent, often focusing on the invariant case, and when general, relying on impractically large intermediate representations.

Prior Work. *Here we summarize, to the best of our knowledge, known universality results, recasting them within a unified framework of universality classes and separation-constrained approximability.*

1. *The classical universality theorem of Pinkus (1999), which states that neural networks can approximate any continuous function, translates in this framework as $\mathcal{U}_\sigma(L, L) = \mathcal{C}(\mathbb{R}, \mathbb{R})$.*

2. *Segol & Lipman (2020) show that a simplified version of 3-layer PointNets, where convolutional filters of width 1 appear only in certain layers (see Examples 1.iii and iv), is universal. This, in turn, implies that full 3-layer PointNets are universal in the class of continuous S_n-equivariant functions. Namely, $\mathcal{U}_\sigma(C, P, C) = \mathcal{U}_\sigma(P, P, P) = \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$.*

3. *Ravanbakhsh (2020) shows that shallow equivariant networks with regular representations as hidden layers are universal among G-equivariant functions. Namely, for permutation representations V and W, $\mathcal{U}_\sigma(M, N) = \mathcal{C}_G(V, W)$ where $M = \mathrm{Aff}_G(V, \mathbb{R}^G)$ and $N = \mathrm{Aff}_G(\mathbb{R}^G, W)$.*

4. *Joshi et al. (2023) show that models expressive enough to distinguish all G-orbits become universal in the invariant sense once augmented with a shallow neural network head. Namely, if the neural space $\mathcal{N}_\sigma(M_1, \ldots, M_d, I)$ separates G-orbits in $\mathbb{R}^n$, then the associated universality class, augmented with a shallow network head, satisfies $\mathcal{U}_\sigma(M_1, \ldots, M_d, I, L) = \mathcal{C}_G(\mathbb{R}^n, \mathbb{R})$.*

5. *Geerts (2020); Maron et al. (2019a); Chen et al. (2019) show that graph neural networks and invariant graph networks (Maron et al., 2018) can approximate any continuous invariant function with the same separation power as the Weisfeiler–Leman test. Namely, for the layer space $M = \mathrm{Aff}_{S_n}((\mathbb{R}^n)^{\otimes k}, (\mathbb{R}^n)^{\otimes k})$, which processes k-order relational structures equivariantly, $\mathcal{U}_\sigma(\underbrace{M, ..., M}_{d \text{ times}}, I, L) = \mathcal{C}_{k\text{-}WL_d}\big((\mathbb{R}^n)^{\otimes k}, (\mathbb{R}^n)^{\otimes k}\big)$, that is, the set of continuous functions with the same separation power as the k-WL test after d iterations.*

6. *Moreover, Pacini et al. (2025) show that in some cases the final trivial layer, as in the two previous examples, is necessary for separation-constrained universality when certain representations are involved. Namely, they prove that $\mathcal{U}_\sigma(C, I) \subsetneq \mathcal{U}_\sigma(P, I) \subsetneq \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R})$, even though these spaces exhibit the same separation power: $\rho\big(\mathcal{U}_\sigma(C, I)\big) = \rho\big(\mathcal{U}_\sigma(P, I)\big) = \rho\big(\mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R})\big)$.*

4 Separation-constrained Universality for Invariant Networks

In this section, we establish a general result on separation-constrained universality (Definition 5) for invariant neural networks, extending prior works on invariant universality (see Prior Work 4 and 5). In particular, we prove that pathological mismatches between separation power and approximation power (see Prior Work 6) can always be resolved by adding a fully connected readout layer.

Theorem 1. *Let $M_1, \ldots, M_d$ be layer spaces as defined in Definition 1 and recall that I denotes the layer space of invariant linear functions from Example 1.ii. Set $\rho = \rho\big(\mathcal{U}_\sigma(M_1, \ldots, M_d, I)\big)$. Then*

$$\mathcal{U}_\sigma(M_1, \ldots, M_d, I, L) = \mathcal{C}_\rho(V). \tag{4}$$

Proof of Theorem 1. Note that by Theorem 4 in Pacini et al. (2024b) and the remark following it, ρ is preserved under the extension from $\mathcal{N}_\sigma(M_1, \ldots, M_d, I, L)$ to $\mathcal{U}_\sigma(M_1, \ldots, M_d, I, L)$ and from $\mathcal{N}_\sigma(M_1, \ldots, M_d, I)$ to $\mathcal{U}_\sigma(M_1, \ldots, M_d, I)$, therefore

$$\rho = \rho(\mathcal{N}_\sigma(M_1, \ldots, M_d, I)) = \rho(\mathcal{N}_\sigma(M_1, \ldots, M_d, I, L)).$$

Hence we get $\mathcal{U}_\sigma(M_1, \ldots, M_d, I, L) \subseteq \mathcal{C}_\rho(V)$ and we only have to prove the opposite inclusion. Given functions $f_1, \ldots, f_h \in \mathcal{C}(V, \mathbb{R})$, define their parallelization as $F_h = (f_1, \ldots, f_h) \colon V \longrightarrow \mathbb{R}^h$, $F_h(x) = (f_1(x), \ldots, f_h(x))$, and set

$$\mathcal{A}_h := \big\{\eta \circ F_h \mid \eta \in \mathcal{C}(\mathbb{R}^h)\big\}, \quad \mathcal{A}'_h := \big\{\eta \circ F_h \mid \eta \in \bigcup_{k \in \mathbb{N}} \mathcal{N}_\sigma(L^{h\times k}, L^{k\times 1})\big\}. \tag{5}$$

Note that by the universal approximation theorem, $\mathcal{A}_h = \overline{\mathcal{A}_h} = \overline{\mathcal{A}'_h}$. From now on we will take $\mathcal{F} = \{f_h\}_{h\in\mathbb{N}}$ to be a family of functions such that $\rho(\mathcal{F}) = \rho$. We get via a result from the appendix that

$$\mathcal{C}_\rho(V) \overset{\text{Lemma 3}}{=} \overline{\bigcup_{h\in\mathbb{N}} \mathcal{A}_h} = \overline{\bigcup_{h\in\mathbb{N}} \mathcal{A}'_h}. \tag{6}$$

Define

$$\mathcal{N}_h := \bigcup_{\vec{k}\in\mathbb{N}^{d+1}} \mathcal{N}_\sigma(M_1^{1\times k_1}, \dots, M_d^{k_{d-1}\times k_d}, I^{k_d\times h}) \quad \text{for each } h \in \mathbb{N}\,.$$

Then we can write

$$\begin{aligned}
\mathcal{U}_\sigma(M_1,\dots,M_d,I,L) &= \overline{\bigcup_{\vec{k}\in\mathbb{N}^{d+1}} \mathcal{N}_\sigma(M_1^{1\times k_1},\dots,M_d^{k_{d-1}\times k_d}, I^{k_d\times k}, L^{k\times 1})} \\
&= \overline{\bigcup_{\vec{k}\in\mathbb{N}^{d+2}} \mathcal{N}_\sigma(M_1^{1\times k_1},\dots,M_d^{k_{d-1}\times k_d}, I^{k_d\times h}) \,\hat{\circ}\, \mathcal{N}_\sigma(L^{h\times k}, L^{k\times 1})} \\
&= \overline{\bigcup_{h\in\mathbb{N}} \left\{ \eta\circ f \mid f\in\mathcal{N}_h,\ \eta\in \bigcup_{k\in\mathbb{N}} \mathcal{N}_\sigma(L^{h\times k}, L^{k\times 1}) \right\}} \\
&\overset{\text{Equation 5}}{\supseteq} \overline{\bigcup_{h\in\mathbb{N}} \mathcal{A}'_h} \overset{\text{Equation 6}}{=} \mathcal{C}_\rho(V).
\end{aligned}$$

To prove the above inclusion, if $f_1,\dots,f_h \in \mathcal{N}_\sigma(M_1,\dots,M_d,I)$ then their parallelization $(f_1,\dots,f_h)$ belongs to $\mathcal{N}_h$ by Lemma 1 from the appendix. The last equality holds because of Equation 6 and Corollary 2, since there exists a family of networks $\mathcal{F} = \{f_h\}_{h\in\mathbb{N}}$ such that $f_h \in \mathcal{N}_\sigma(M_1,\dots,M_d,I)$ for each $h\in\mathbb{N}$ and $\rho(\mathcal{F}) = \rho$, and we can use this family to define $\mathcal{A}'_h$. □

5 Universality of Equivariant Neural Networks

In this section, we extend the previous results to the equivariant setting. However, important differences between the invariant and equivariant cases emerge: in Section 5.1 we show that the standard form of the separation relation as in Definition 5 often fails to faithfully characterize equivariant universality classes, requiring us to introduce the notion of entry-wise separation (Definition 6). In Section 5.2, we establish universality theorems analogous to Theorem 1, showing that the outcome crucially depends on the choice of output space.

5.1 Entry-wise Separation

Here, we study equivariant functions by reducing the problem to the analysis of suitable invariant functions, thereby connecting our setting to the results of Section 4. The main tool for this reduction is the projection onto output coordinates. More precisely, let G be a finite group acting on the finite set X. For $x\in X$ consider the stabilizer of x, given by $G_x = \mathrm{Stab}_G(x) := \{g\in G \mid gx = x\}$, and the linear projection $\pi_x : \mathbb{R}^X \to \mathbb{R}$ onto the x-th coordinate. Then π_x induces the pushforward map

$$\begin{aligned}
\pi_{x*}\colon \mathcal{C}_G(V,\mathbb{R}^X) &\longrightarrow \mathcal{C}_{G_x}(V) \\
f &\longmapsto \pi_x\circ f.
\end{aligned}$$

Since the vector of projections satisfies $(\pi_x)_{x\in X} = \mathrm{id}_{\mathbb{R}^X}$, it follows that $(\pi_{x*})_{x\in X}$ acts as the identity on $\mathcal{C}_G(V,\mathbb{R}^X)$. Thus, the study of universality for equivariant maps reduces to the problem of synchronous universality of the invariant projection maps. However, below Proposition 1 shows that the interaction between equivariance and the global separation ρ is non-trivial when projecting functions onto different output entries.

Proposition 1. *Let $\rho = \rho(\mathcal{N})$ be the separation relation of a family of equivariant neural networks $\mathcal{N}$. The restriction of π_x to*

$$\mathcal{C}_{G,\rho}(V,\mathbb{R}^X) := \mathcal{C}_G(V,\mathbb{R}^X) \cap \mathcal{C}_\rho(V,\mathbb{R}^X)$$

is surjective onto $\mathcal{C}_{G_x,\rho}(V)$, the space of G_x-invariant functions with separation relation ρ.

The proof for Proposition 1 and of all subsequent results may be found in the Appendix.

Proposition 1 shows that, after projection onto a single output coordinate, the space of equivariant functions with separation ρ is constrained by a stricter relation. This relation combines ρ with the

G_x-invariance relation, which identifies elements within each G_x-orbit. However, the following example shows that this stricter condition remains insufficient to correctly characterize the universality classes associated with equivariant architectures.

Example 3 (Separation for CNNs)**.** *Let C be the layer space of convolutional filters with width 1 defined in Example 1.iii. For the purpose of this example, it is sufficient to restrict C to go from $\mathbb{R}^n$ to $\mathbb{R}^n$ with S_n acting in the standard way on $\mathbb{R}^n$. Hence, (2) becomes*

$$C := \left\{ v \mapsto x\,\mathrm{id}\cdot v + y\mathbb{1} \;\middle|\; x, y \in \mathbb{R} \right\}.$$

Consider the universality class for $d \geq 2$:

$$\mathcal{U}_\sigma^d(C) := \mathcal{U}_\sigma(\underbrace{C, \ldots, C}_{d \text{ times}}).$$

We can show (see Proposition 5) that

$$\mathcal{U}_\sigma^d(C) = \{(x_1, \ldots, x_n) \mapsto (f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{C}(\mathbb{R})\} \subsetneq \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n). \tag{7}$$

Note that $\mathrm{id}_{\mathbb{R}^n} \in \mathcal{U}_\sigma^d(C)$. Then, $\rho(\mathcal{U}_\sigma^d(C))$ is the trivial separation relation, namely $\rho(\mathcal{U}_\sigma^d(C)) = \{(x, x) \mid x \in \mathbb{R}^n\}$. Thus, the target space of separation-constrained universality is $\mathcal{C}_{S_n,\rho}(\mathbb{R}^n, \mathbb{R}^n) = \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$. However, (7) shows that $\mathcal{U}_\sigma^d(C) \subsetneq \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$ for each $d \geq 2$. Or equivalently, in this case separation-constrained universality can never be attained, regardless of depth d.

Example 3 shows that characterizing equivariant universality classes requires a finer notion of separability, which we now define.

Definition 6 (Entry-wise Separation)**.** *Let G be a finite group acting on a finite set $X = \{x_1, \ldots, x_n\}$, and let $\mathbb{R}^X$ denote the associated permutation representation. Let V be another permutation representation over G and $\mathcal{N}$ a neural space of functions in $\mathcal{C}_G(V, \mathbb{R}^X)$. Let $\pi_x : \mathbb{R}^X \to \mathbb{R}$ be the linear projection onto the x-th component for each $x \in X$. Define the family of separation relations*

$$\rho_x(\mathcal{N}) := \{(\alpha, \beta) \in V \times V \mid \pi_x f(\alpha) = \pi_x f(\beta) \text{ for all } f \in \mathcal{N}\}.$$

for each $x \in X$. We define the entry-wise separation relation *as the collection of separation relations*

$$\overline{\rho}(\mathcal{N}) = \big(\rho_{x_1}(\mathcal{N}), \ldots, \rho_{x_n}(\mathcal{N})\big).$$

We define the set of continuous functions that respect $\overline{\rho}$ as

$$\mathcal{C}_{\overline{\rho}}(V, \mathbb{R}^X) := \big\{ f \in \mathcal{C}(V, \mathbb{R}^X) \mid \pi_x f(v_1) = \pi_x f(v_2)\ \forall (v_1, v_2) \in \rho_x(\mathcal{N}), \forall x \in X \big\}.$$

If a universality class with entry-wise separation $\overline{\rho}$ coincides with $\mathcal{C}_{\overline{\rho}}$, we call it ***entry-wise separation universal****.*

Note that $\rho(\mathcal{N}) = \rho_{x_1}(\mathcal{N}) \cap \cdots \cap \rho_{x_n}(\mathcal{N})$, so the standard separation relation is implied by the entry-wise separation relations. That is $\mathcal{N} \subseteq \mathcal{C}_{\overline{\rho}(\mathcal{N})}(V, \mathbb{R}^X) \subseteq \mathcal{C}_{\rho(\mathcal{N})}(V, \mathbb{R}^X)$. As noted in Section 3.3, separation is a necessary condition for approximation, and now we see entry-wise separation is necessary as well. Note that in certain cases entry-wise separation reduces entirely to the standard separation relation, for instance in the invariant case where G acts trivially on $\mathbb{R}$, or more simply when $\rho(\mathcal{N}) = \rho_{x_1}(\mathcal{N}) = \cdots = \rho_{x_n}(\mathcal{N})$. Yet, Example 3 shows that entry-wise separation can, in fact, be strictly stronger than standard separation. Indeed, on the one hand (7) gives

$$\pi_{1*}\mathcal{U}_\sigma^d(C) = \{\, (x_1, \ldots, x_n) \mapsto f(x_1) \mid f \in \mathcal{C}(\mathbb{R}) \,\} \subsetneq \mathcal{C}_{\mathrm{Stab}_{S_n}(1)}(\mathbb{R}^n, \mathbb{R}),$$

while on the other hand, we have $\pi_{1*}\mathcal{U}_\sigma^d(C) = \mathcal{C}_{\rho_1}(\mathbb{R}^n, \mathbb{R})$. If we denote $\mathbb{R}^n = \mathbb{R} \times \mathbb{R}^{n-1}$, here $\rho_1 := \big\{ ((x_1, \overline{x}), (y_1, \overline{y})) \in (\mathbb{R} \times \mathbb{R}^{n-1})^2 \;\big|\; x_1 = y_1 \big\}$. Analogous results hold for the other ρ_i, with $i = 2, \ldots, n$. This proves the following proposition and shows that the universality class in Example 3 can be completely characterized by entry-wise separation universality.

Proposition 2. *Define $\overline{\rho} = \overline{\rho}\big(\mathcal{U}_\sigma^d(C)\big)$. Then, $\mathcal{U}_\sigma^d(C) = \mathcal{C}_{\overline{\rho}}(\mathbb{R}^n, \mathbb{R}^n)$.*

5.2 Entry-wise Separation Constrained Universality

Now we are ready to state universality results under the more general notion of entry-wise separability as discussed in Section 5.1.

Theorem 2. *Let $V_0, \dots, V_h$ be permutation representations of a finite group G. Let X be a finite G-set and $\mathbb{R}^X$ its associated permutation representation. Let $M_1, \dots, M_f$ be layer spaces in $\mathrm{Aff}_G(V_{i-1}, V_i)$ for $i = 1, \dots, f$, and let M be a layer space in $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X)$ containing the identity map. Let d be such that*

$$\overline{\rho} := \overline{\rho}\big(\mathcal{U}_\sigma(M_1, \dots, M_f, \underbrace{M, ..., M}_{d \text{ times}})\big) = \overline{\rho}\big(\mathcal{U}_\sigma(M_1, \dots, M_f, \underbrace{M, ..., M}_{d+1 \text{ times}})\big). \tag{8}$$

Then,

$$\mathcal{U}_\sigma(M_1, \dots, M_f, \underbrace{M, ..., M}_{d+2 \text{ times}}) = \mathcal{C}_{\overline{\rho}}(V_0, \mathbb{R}^X).$$

In other words, repeating the output layer beyond the separation-stabilization threshold ensures entry-wise separation-constrained universality.

Since by Theorem 3 of Pacini et al. (2024b), separation is known to stabilize after a certain depth, we obtain the following corollary.

Corollary 1. *Assume the notation of Theorem 2. There exists a natural number D for which $\mathcal{U}_\sigma(M_1, \dots, M_f, \underbrace{M, ..., M}_{d \text{ times}})$ is entry-wise separation-constrained universal for each $d \geq D$.*

In a different direction, we can show that entry-wise separation-constrained universality can be achieved when the output layer is a convolutional filter of width 1, without the requirement of sufficient depth as in Theorem 2 and Corollary 1. This is formalized as follows.

Theorem 3. *Let $V_0, \dots, V_f$ be permutation representations of a finite group G. Let X be a finite G-set and $\mathbb{R}^X$ its associated permutation representation. Let $M_1, \dots, M_f$ be layer spaces in $\mathrm{Aff}_G(V_{i-1}, V_i)$ for $i = 1, \dots, f$, and let C be a layer space in $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X)$ of convolutional filters with width 1 as defined in Example 1.iii. Then $\mathcal{U}_\sigma(M_1, \dots, M_f, C) = \mathcal{C}_{\overline{\rho}}(V)$, where $\overline{\rho} := \overline{\rho}\big(\mathcal{U}_\sigma(M_1, \dots, M_f, C)\big)$.*

Note that when C is defined on a one-dimensional space, we have $C = L$, and M^d becomes the invariant layer space I. In this case, Theorem 3, which is formulated in the equivariant setting, specializes to Theorem 1—the corresponding result in the invariant setting.

At first sight, it may be tempting to compare Theorem 2 and Theorem 3 and conclude that Theorem 3 is a stronger statement. However, it is important to note that adding the C layer space at the end does not change the entry-wise separation power of the model class, whereas adding a certain number of M layers may increase it. Theorem 2 explicitly accounts for this effect.

Theorem 2 and Corollary 1 may be particularly relevant for their practical implications: they ensure that maximal expressivity is reached at finite depth and rule out the possibility of unbounded improvement. Theorem 3, on the other hand, is instrumental in recovering known results such as (Segol & Lipman, 2020). It also shows that universality stabilization in Theorem 2 and Corollary 1 can occur at the same depth as entry-wise separation stabilization, revealing that the threshold in Theorem 2 is not always optimal.

Remark 1. Thanks to Theorem 3, we can easily recover the universality result of Segol & Lipman (2020). Namely, $\mathcal{U}_\sigma(C, P, C) = \mathcal{U}_\sigma(P, P, P) = \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$. It remains to verify that $\pi_i^* \mathcal{N}_\sigma(C, P, C)$ separates $\mathrm{Stab}_{S_n}(i)$-orbits in $\mathbb{R}^n$, which follows directly from Lemma 5 (Appendix C.2).

Note that this shows that the depth threshold required for separation-stability in Theorem 2 provides a *sufficient*, but not *necessary*, condition for universality. Indeed, $\overline{\rho}\big(\mathcal{U}_\sigma(P, P, P)\big) \subsetneq \overline{\rho}\big(\mathcal{U}_\sigma(P, P)\big)$, so separation has not stabilized, yet entry-wise separation universality is already achieved. However, determining in general when separation stabilization takes place is a difficult problem. Corollary 1 guarantees that maximal expressivity is reached after a finite number of steps and then saturates. This result supports the intuition that increasing depth enhances expressivity. Less intuitively, it also shows

that beyond a certain threshold, saturation occurs and further increases in depth no longer affect the universality class.

Remark 2. Theorem 3 marks a significant difference between the equivariant and the invariant cases. Indeed, Pacini et al. (2025) shows that, although $\rho(\mathcal{U}_\sigma(C, I)) = \rho(\mathcal{U}_\sigma(P, I)) = \rho(\mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}))$, the corresponding universality classes satisfy $\mathcal{U}_\sigma(C, I) \subsetneq \mathcal{U}_\sigma(P, I) \subsetneq \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R})$. These strict inequalities are proved via a characterization through differential operators. In the equivariant case, we have $\mathcal{U}_\sigma(C, C) \subsetneq \mathcal{U}_\sigma(P, C) \subseteq \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n)$, yet as we showed here, both spaces can be characterized in terms of entry-wise separation, without resorting to the differential operator characterization. Note that we expect this to be a phenomenon specific to networks with output layers in $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X)$. Output spaces in $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R})$, or more generally in $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^Y)$, may instead require a characterization in terms of differential operators for arbitrary finite G-sets Y.

6 Limitations

Our results characterize universality of deep invariant and equivariant networks under separation constraints, but several limitations remain which provide avenues for future work. First, the theory applies to networks with point-wise activations and permutation representations. Extending the analysis to other types of representations or more general nonlinearities may require different approaches. Second, our universality theorems are asymptotic and do not provide quantitative approximation rates or sample complexity bounds, which are important for understanding expressivity in practice. Finally, we have not addressed optimization or trainability: while depth is shown to be sufficient for universality, when such networks can be efficiently trained remains an open question.

7 Conclusions

We established new universality results for deep invariant and equivariant networks. For invariance, we proved that depth is sufficient to guarantee universality within the class of separation-respecting functions. For equivariance, we introduced the refined concept of entry-wise separability and showed that, once entry-wise relations stabilize, deep equivariant networks achieve universality. Taken together, these results unify and extend prior shallow or architecture-specific universality theorems, highlighting depth as a general mechanism for universality in equivariant models. We hope this framework provides a basis for future advances in the analysis and design of expressive, symmetry-aware, neural networks.

Reproducibility Statement

This work is purely theoretical and contains no experiments or datasets. All results are formally stated as theorems, propositions, or corollaries, and complete proofs are provided in the main text and appendices. Definitions and assumptions are explicitly stated to ensure mathematical clarity, and we reference relevant prior results where appropriate. As such, all claims in the paper can be fully verified by checking the provided proofs.

Ethics Statement

This work is theoretical and does not involve experiments with human subjects, sensitive data, or deployment of models in real-world applications. We therefore do not foresee any direct ethical concern.

References

Claudio Battiloro, Ege Karaismailoğlu, Mauricio Tec, George Dasoulas, Michelle Audirac, and Francesca Dominici. E(n) Equivariant Topological Neural Networks, February 2025. URL http://arxiv.org/abs/2405.15429. arXiv:2405.15429 [cs].

Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, and Risi Kondor. Lorentz Group Equivariant Neural Network for Particle Physics. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 992–1002. PMLR, November 2020. URL https://proceedings.mlr.press/v119/bogatskiy20a.html. ISSN: 2640-3498.

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.

Ovidiu Calin. *Deep Learning Architectures: A Mathematical Approach*. Springer Publishing Company, Incorporated, 1st edition, 2020. ISBN 978-3-030-36720-6.

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].

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

Nadav Dym and Steven J. Gortler. Low Dimensional Invariant Embeddings for Universal Geometric Learning, May 2022. URL http://arxiv.org/abs/2205.02956. arXiv:2205.02956 [cs, math].

Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual Pathway Priors for Soft Equivariance Constraints. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30037–30049. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/fc394e9935fbd62c8aedc372464e1965-Abstract.html.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html?utm_source=chatgpt.com.

Floris Geerts. The expressive power of kth-order invariant graph networks, July 2020. URL http://arxiv.org/abs/2007.12035. arXiv:2007.12035 [cs, math, stat].

Floris Geerts and Juan L Reutter. EXPRESSIVENESS AND APPROXIMATION PROPERTIES OF GRAPH NEURAL NETWORKS. pp. 43, 2022.

Haojie Huang, Owen Lewis Howell, Dian Wang, Xupeng Zhu, Robert Platt, and Robin Walters. Fourier Transporter: Bi-Equivariant Robotic Manipulation in 3D. October 2023. URL https://openreview.net/forum?id=UulwvAU1W0.

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_.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

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.

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*, pp. 2747–2755. PMLR, July 2018. URL https://proceedings.mlr.press/v80/kondor18a.html. ISSN: 2640-3498.

Maxime W. Lafarge, Erik J. Bekkers, Josien P. W. Pluim, Remco Duits, and Mitko Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis*, 68:101849, February 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101849. URL https://www.sciencedirect.com/science/article/pii/S1361841520302139.

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.

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

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*, pp. 4363–4371. PMLR, May 2019b. URL https://proceedings.mlr.press/v97/maron19a.html. ISSN: 2640-3498.

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.

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

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

Marco Pacini, Gabriele Santin, Bruno Lepri, and Shubhendu Trivedi. On Universality Classes of Equivariant Networks, June 2025. URL http://arxiv.org/abs/2506.02293. arXiv:2506.02293 [cs].

Mircea Petrache and Shubhendu Trivedi. Approximation-Generalization Trade-offs under (Approximate) Group Equivariance. *Advances in Neural Information Processing Systems*, 36:61936–61959, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c35f8e2fc6d81f195009a1d2ae5f6ae9-Abstract-Conference.html.

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.

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)*, pp. 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/.

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

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

Jean-Pierre Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1977. ISBN 978-1-4684-9460-0 978-1-4684-9458-7. doi: 10.1007/978-1-4684-9458-7. URL http://link.springer.com/10.1007/978-1-4684-9458-7.

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, pp. 38680–38694, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-71387-108-8.

Matus Telgarsky. Benefits of depth in neural networks, May 2016. URL http://arxiv.org/abs/1602.04485. arXiv:1602.04485 [cs, stat].

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].

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9323–9332. PMLR, July 2021. URL https://proceedings.mlr.press/v139/satorras21a.html. ISSN: 2640-3498.

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.

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].

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.

A PRELIMINARIES

Definition 7 (Group). *A* group *is a set G together with a binary operation $\cdot : G \times G \to G$ such that:*

- ***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 finite *if G has finitely many elements.*

We next recall the notion of a group homomorphism, which is a structure-preserving map between groups.

Definition 8 (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 satisfies*

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

Definition 9 (Group Actions). *Let G be a group and let X be a set. A* group action *of G on X is a map*

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

often written as $\phi_g(x) = \Phi(g, x)$ for $g \in G$ and $x \in X$, such that:

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

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

A set X together with a group action of G is called a G-set. *Equivalently, X is a G-set if there exists an action $\cdot : G \times X \to X$ satisfying the identity and compatibility conditions above.*

A central notion in our analysis is that of a map between G-sets that respects the group action, leading to the following definition.

Definition 10 (Equivariance). *Let X and Y be G-sets. A function $f : X \to Y$ is* G-equivariant *if, for all $g \in G$ and $x \in X$, we have*

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

Definition 11 (Group Representations). *Let G be a group and let V be a vector space over $\mathbb{R}$. A* representation *of G on V is a group homomorphism*

$$\phi : G \to \mathrm{GL}(V),$$

where $\mathrm{GL}(V)$ denotes the group of invertible linear maps $V \to V$. Given such a homomorphism, we obtain a linear action of G on V by setting

$$gv \coloneqq \phi(g)(v) \quad \text{for } g \in G,\ v \in V.$$

When V and W are G-representations, we denote by $\mathrm{Hom}_G(V, W)$ the space of G-equivariant linear maps $V \to W$, and by $\mathrm{Aff}_G(V, W)$ the set of G-equivariant affine maps $V \to W$.

Note that $\mathrm{Hom}_G(V, W)$ is a vector space. Indeed, $0 \in \mathrm{Hom}_G(V, W)$, and for any $f, g \in \mathrm{Hom}_G(V, W)$ and $\alpha, \beta \in \mathbb{R}$, the linear combination $\alpha f + \beta g$ is still in $\mathrm{Hom}_G(V, W)$. The same property holds for $\mathrm{Aff}_G(V, W)$.

Definition 12 (Permutation Representations). *Let X be a finite set and let G be a finite group acting on X. The associated* permutation representation *of G is the linear action of G on $\mathbb{R}^X$ defined on the standard basis $\{e_x\}_{x \in X}$ by*

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

We now make explicit how this fits the general notion of a representation.

Proposition 3. *Let X and G be as above and set $V \coloneqq \mathbb{R}^X$. For each $g \in G$ there is a unique linear map*

$$\phi(g) : V \to V$$

such that $\phi(g)(e_x) = e_{g\cdot x}$ for all $x \in X$. Then the map

$$\phi : G \to \mathrm{GL}(V), \quad g \mapsto \phi(g)$$

is a representation of G on V.

Proof. Since $\{e_x\}_{x\in X}$ is a basis of $V = \mathbb{R}^X$, for each $g \in G$ there exists a unique linear map $\phi(g) : V \to V$ such that $\phi(g)(e_x) = e_{g\cdot x}$ for all $x \in X$. Moreover,

$$\phi(g^{-1})(\phi(g)(e_x)) = \phi(g^{-1})(e_{g\cdot x}) = e_{g^{-1}\cdot(g\cdot x)} = e_x,$$

so $\phi(g^{-1})$ is the inverse of $\phi(g)$ and $\phi(g) \in \mathrm{GL}(V)$.

For $g, h \in G$ and any $x \in X$ we have

$$(\phi(g)\phi(h))(e_x) = \phi(g)(e_{h\cdot x}) = e_{g\cdot(h\cdot x)} = e_{(gh)\cdot x} = \phi(gh)(e_x),$$

hence $\phi(g)\phi(h) = \phi(gh)$ and $\phi : G \to \mathrm{GL}(V)$ is a group homomorphism. □

Moreover, after choosing an ordering $X = \{x_1, \dots, x_n\}$ and identifying $V \cong \mathbb{R}^n$, each $\phi(g)$ is represented by a permutation matrix $P_g \in \mathbb{R}^{n\times n}$ with entries

$$(P_g)_{ij} = 1 \text{ if } g \cdot x_j = x_i, \text{ and } (P_g)_{ij} = 0 \text{ otherwise.}$$

In particular, $P_{gh} = P_g P_h$ and $P_e = I_n$, so $g \mapsto P_g$ is a group homomorphism into $\mathrm{GL}_n(\mathbb{R})$.

B Separation-constrained Universality for Invariant Networks

We recall the notation introduced in equation 3 for the subspace M of $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^Y)$.

$$M^{k\times h} := \left\{ \begin{array}{c} (x_1, \dots, x_k) \mapsto \left(\sum_{j=1}^k f_{1,j}(x_j), \dots, \sum_{j=1}^k f_{h,j}(x_j)\right) \\ f_{ij} \in M, i = 1, \dots, h, j = 1, \dots, k \end{array} \right\}.$$

Thus, $M^{k\times h} \subseteq \mathrm{Aff}_G(\mathbb{R}^X \otimes \mathbb{R}^k, \mathbb{R}^Y \otimes \mathbb{R}^h)$. In particular, for each f_{ij} in the above definition, we write

$$f_{ij}(x) = A_{ij}x + b_{ij},$$

where A_{ij} and b_{ij} denote respectively the linear part and the translational part of f_{ij}. With this notation, the linear and translational parts of an element in $M^{k\times h}$ can be written respectively as

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ A_{21} & A_{22} & \cdots & A_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{h1} & A_{h2} & \cdots & A_{hk} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \sum_{j=1}^k b_{1j} \\ \sum_{j=1}^k b_{2j} \\ \vdots \\ \sum_{j=1}^k b_{hj} \end{bmatrix}.$$

The following definitions and lemmas are used for the proof of Theorem 1.

Lemma 1. *Let $f_1, \dots, f_h \in \mathcal{N}_\sigma(M_1, \dots, M_d, I)$ then their parallelization $(f_1, \dots, f_h)$ belongs to $\bigcup_{\tilde{k}\in\mathbb{N}^{d+1}} \mathcal{N}_\sigma(M_1^{1\times k_1}, \dots, M_d^{k_{d-1}\times h})$.*

Proof of Lemma 1. Let us consider the case $h = 2$; for $h > 2$, the proof is analogous.

Let $f_1, f_2 \in \mathcal{N}_\sigma(M_1, \dots, M_d)$. Then each affine layer at depth i in (f_1, f_2) is a block diagonal matrix whose first block is the i-th layer of f_1 and the second is the i-th layer of f_2; a similar analysis holds for the bias terms. Hence, this layer belongs to $M_i^{2\times 2}$ for $i > 1$. For $i = 1$, the first linear layer of (f_1, f_2) is a block column matrix where each block is the first layer of f_1 and f_2; again, a similar analysis holds for the bias terms. Hence, this layer belongs to $M_1^{1\times 2}$. This shows that

$$(f_1, f_2) \in \mathcal{N}_\sigma(M_1^{1\times 2}, M_2^{2\times 2}, \dots, M_d^{2\times 2}) \subseteq \bigcup_{\tilde{k}\in\mathbb{N}^{d-1}} \mathcal{N}_\sigma(M_1^{1\times k_1}, \dots, M_d^{k_{d-1}\times 2}).$$

□

Definition 13. *Let $M_1, \dots, M_d$ be layer spaces. Let $\mathcal{B}_i$ be bases for the layer space M_i, and define*

$$M_i^{\mathbb{Q}} := \mathrm{Span}_{\mathbb{Q}}\, \mathcal{B}_i$$

for each $i = 1, \dots, d$. Define rational neural spaces as follows:

$$\mathcal{N}_\sigma{}^{\mathbb{Q}}(M_1, \dots, M_d) := \mathcal{N}_\sigma(M_1^{\mathbb{Q}}, \dots, M_d^{\mathbb{Q}}).$$

Note that $\mathcal{N}_\sigma{}^{\mathbb{Q}}(M_1, \dots, M_d)$ depends *on the choice of the bases $\mathcal{B}_1, \dots, \mathcal{B}_d$.*

Lemma 2. *In the notation of Definition 13,*

$$\rho(\mathcal{N}_\sigma(M_1, \dots, M_d)) = \rho(\mathcal{N}_\sigma{}^{\mathbb{Q}}(M_1, \dots, M_d)).$$

Therefore, $\rho(\mathcal{N}_\sigma{}^{\mathbb{Q}}(M_1, \dots, M_d))$ does not depend on the choice of bases $\mathcal{B}_1, \dots, \mathcal{B}_d$.

Proof of Lemma 2. By the continuity of the parametrization map and the density of $M_i^{\mathbb{Q}}$ in M_i. □

Lemma 2 implies the following corollary.

Corollary 2. *There exists a countable family $\mathcal{F} = \{f_h\}_{h \in \mathbb{N}} \subseteq \mathcal{N}_\sigma(M_1, \dots, M_d)$ such that*

$$\rho(\mathcal{F}) = \rho(\mathcal{N}_\sigma(M_1, \dots, M_d)).$$

Lemma 3. *Let $V = \mathbb{R}^d$ with its usual topology and let ρ be a* closed *equivalence relation on V. For a family $\mathcal{F} = \{f_n\}_{n \in \mathbb{N}}$ of continuous maps $f_n : V \to \mathbb{R}^m$ such that $\rho(\mathcal{F}) = \rho$. Then the set*

$$\mathcal{A} := \bigcup_{n \geq 1} \left\{ A(f_1, \dots, f_n)|_K \;:\; A \in \mathcal{C}\left((\mathbb{R}^m)^n, \mathbb{R}\right) \right\}$$

is dense in $\mathcal{C}_\rho(V)$. Or equivalently, for every $h \in \mathcal{C}_\rho(V)$ there exist $n_k \uparrow \infty$ and $A_{n_k} \in \mathcal{C}((\mathbb{R}^m)^{n_k}, \mathbb{R})$ such that $A_{n_k}(f_1, \dots, f_{n_k}) \to h$.

Proof of Lemma 3. For $x \in V$ set $\widehat{F}(x) := (f_n(x))_{n \in \mathbb{N}}$. Fix a compact $K \subset V$. Since each f_n is continuous, $\widehat{F}(K)$ is compact in the product $V^{\mathbb{N}}$. Note that $\rho = \{(x, y) \in V^2 : \widehat{F}(x) = \widehat{F}(y)\}$, so that the map $\phi : K/\rho \to \widehat{F}(K)$ defined by $\phi([x]) := \widehat{F}(x)$ is well defined. Furthermore ϕ is continuous and bijective, and since K/ρ is compact Hausdorff and $\widehat{F}(K)$ is Hausdorff, ϕ is a homeomorphism. Hence every $h \in \mathcal{C}_\rho(K)$ factors uniquely as

$$h = H \circ \widehat{F}|_K \qquad \text{for a unique } H \in \mathcal{C}(\widehat{F}(K), \mathbb{R}).$$

Let $\pi_n : (\mathbb{R}^m)^{\mathbb{N}} \to (\mathbb{R}^m)^n$ be the projection onto the first n coordinates. Note that

$$\mathcal{A} = \bigcup_{n \geq 1} \left\{ A \circ \pi_n\big|_{\widehat{F}(K)} \;:\; A \in \mathcal{C}\left((\mathbb{R}^m)^n, \mathbb{R}\right) \right\}.$$

Then $\mathcal{A}$ is a sub-algebra of $\mathcal{C}(\widehat{F}(K))$ containing constants. We next prove that $\mathcal{A}$ separates points. Indeed, if $y, y' \in \widehat{F}(K)$ with $y \neq y'$, then there exists j with $y_j \neq y'_j$; define the continuous scalar function $p : \mathbb{R}^m \to \mathbb{R}$ such as $p(u) = \langle u, y_j - y'_j \rangle$, and note that $p(y_j) \neq p(y'_j)$ and choose $A \in \mathcal{C}((\mathbb{R}^m)^n)$ given by $A(z_1, \dots, z_j, \dots) := p(z_j)$. This function A lies in $\mathcal{A}$ and satisfies $(A \circ \pi_j)(y) \neq (A \circ \pi_j)(y')$ as desired. Now we may use the Stone–Weierstrass theorem, which gives that $\overline{\mathcal{A}} = \mathcal{C}(\widehat{F}(K))$ in the uniform norm, concluding the proof. □

C Universality of Equivariant Neural Networks

C.1 Entry-wise Separation

In this section, we study equivariant functions by reducing the problem to the analysis of particular invariant functions, thereby extending the previous results. The tools used for this reduction are suitable projections onto output coordinates, together with reconstruction maps that allow us to

recover the entire function from a single projection. For the sake of presentation, we begin by considering the case where G acts transitively on X. Let $x \in X$, and let G_x denote the stabilizer of x. Let $\pi_x : \mathbb{R}^X \to \mathbb{R}$ be the linear projection on the x-th coordinate in $\mathbb{R}^X$. We obtain the pushforward map of π_x, defined as

$$\pi_{x*} : \begin{array}{c} \mathcal{C}_G(V, \mathbb{R}^X) \to \mathcal{C}_{G_x}(V) \\ f \mapsto \pi_x \circ f. \end{array}$$

Let $g_1, \dots, g_t$ be a transversal for G/G_x, that is, a choice of representatives for classes in G/G_x. We define the *reconstruction map* as

$$\theta_x^* : \begin{array}{c} \mathcal{C}_{G_x}(V) \to \mathcal{C}_G(V, \mathbb{R}^X) \\ f \mapsto \left[v \mapsto \sum_{i=1}^t f(g_i^{-1} v) e_{g_i x} \right]. \end{array}$$

Proposition 4. *If the action of G on X is transitive, then reconstruction map θ_x^* is a well-defined, continuous linear operator such that*

(i) $\pi_{x*} \circ \theta_x^* = id_{\mathcal{C}_{G_x}(V)}$,

(ii) $\theta_x^* \circ \pi_{x*} = id_{\mathcal{C}_G(V, \mathbb{R}^X)}$.

Proof. Choosing a different representative for each g_i means choosing an element $g_i \cdot h$ for an arbitrary $h \in G_x$. For $f \in \mathcal{C}_{G_x}(V)$, the G_x-invariance of f implies

$$f((g_i \cdot h)^{-1} v) = f(h^{-1} \cdot g_i^{-1} v) = f(g_i^{-1} v).$$

Then $e_{g_i h x} = e_{g_i x}$ since $h \in G_x$. As a consequence, the choice of representatives $g_1, \dots, g_t$ does not affect $\theta_x^*(f)$. Next, we prove that $\theta_x^*(f)$ is G-equivariant: indeed, for $g \in G$ we have

$$\theta_x^*(f)(gv) = \sum_{i=1}^t f(g_i^{-1} g v) e_{g_i x} = \sum_{i=1}^t f(g_i^{-1} v) e_{g^{-1} g_i x} = g \cdot \sum_{i=1}^t f(g_i^{-1} v) e_{g_i x} = g \cdot \theta_x^*(f)(v),$$

where in the second equality we use the fact that $g^{-1} g_1, \dots, g^{-1} g_t$ is another transversal for G/G_x. These observations prove that θ_x^* is well-defined. It is continuous and linear since it is the composition of continuous and linear functions. We can choose $g_1 = e$, in which case the x-th coefficient in $\theta_x^*(f)(v)$ is simply $f(v)$, proving (i). To prove (ii), notice that for $x \in X$, the set $g_1, \dots, g_t$ is a transversal of G/G_x if and only if $g_1 x, \dots, g_t x$ is the G-orbit of x. Thus for each $f \in \mathcal{C}(V, \mathbb{R}^X)$ we can write

$$f(v) = \sum_{i=1}^t \pi_{g_i x} f(v) e_{g_i x}. \tag{9}$$

Now for $f \in \mathcal{C}_G(V, \mathbb{R}^X)$, we can conclude (ii) as follows:

$$\theta_x^* \pi_{x^*} f(v) = \sum_{i=1}^t \pi_x f(g_i^{-1} v) e_{g_i x} = \sum_{i=1}^t \pi_x g_i^{-1} \cdot f(v) e_{g_i x} = \sum_{i=1}^t \pi_{g_i x} f(v) e_{g_i x} \overset{Equation\ 9}{=} f(v).$$

□

Proposition 4 says that π_{x^*} is a linear homeomorphism, hence, a function class $\mathcal{N}$ is dense in $\mathcal{C}_G(V, \mathbb{R}^X)$ if and only if $\pi_{x^*}(\mathcal{N})$ is. This means that we can restrict ourselves to the study of function families of type $\pi_{x^*}(\mathcal{N})$, which are similar to the study conducted in Section 3.3.

Proof of Proposition 1. The claim follows directly from Proposition 4: applying π_x yields one inclusion, while the reconstruction map yields the other. □

Remark 3 (Linear case). In particular, in the affine case we obtain,

$$\pi_{x*} : \begin{array}{c} \mathrm{Aff}_G(V, \mathbb{R}^X) \to \mathrm{Aff}_{G_x}(V, \mathbb{R}) \\ f \mapsto \pi_x \circ f. \end{array}$$

Note that characterizing $\mathrm{Aff}_{G_x}(V, \mathbb{R})$ reduces to computing V^{G_x}. If $V = \mathbb{R}^Y$ for a finite G-set Y, then we just need to compute the orbits of G_x on Y.

The previous observations translate with minor modifications to the non-transitive case, which we address in the next paragraph. Let $X = Y_1 \sqcup \cdots \sqcup Y_s$ denote the decomposition of X into G-orbits. Observe that

$$\mathcal{C}_G(V, \mathbb{R}^X) = \mathcal{C}_G(V, \mathbb{R}^{Y_1}) \oplus \cdots \oplus \mathcal{C}_G(V, \mathbb{R}^{Y_s}),$$

since the orbits form a disjoint partition of X. Let $x_1, \ldots, x_s$ be elements chosen in $Y_1, \ldots, Y_s$, respectively. Note that

$$\pi_{x_i^*} \mathcal{C}_G(V, \mathbb{R}^X) = \pi_{x_i^*} \mathcal{C}_G(V, \mathbb{R}^{Y_i})$$

for each $i = 1, \ldots, s$. Moreover, for each $i = 1, \ldots, s$, we have

$$\mathcal{C}_G(V, \mathbb{R}^{Y_i}) = \theta^*_{x_i} \pi_{x_i^*} \mathcal{C}_G(V, \mathbb{R}^{Y_i}) = \theta^*_{x_i} \pi_{x_i^*} \mathcal{C}_G(V, \mathbb{R}^X).$$

Thus,

$$\mathcal{C}_G(V, \mathbb{R}^X) = \theta^*_{x_1}\big(\mathcal{C}_{G_{x_1}}(V, \mathbb{R})\big) \oplus \cdots \oplus \theta^*_{x_s}\big(\mathcal{C}_{G_{x_s}}(V, \mathbb{R})\big). \tag{10}$$

In particular, we focus on closed linear subspaces $\mathcal{U} \subseteq \mathcal{C}_G(V, \mathbb{R}^X)$. Equation 10 allows us to restrict our attention to the subspaces $\pi_{x_i^*}\mathcal{U}$, for each $i = 1, \ldots, s$.

Proposition 5. *The following equality is true:*

$$\mathcal{U}_\sigma(\underbrace{C, \ldots, C}_{d \text{ times}}) = \{(x_1, \ldots, x_n) \mapsto (f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{C}(\mathbb{R})\} \subsetneq \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n). \tag{11}$$

Proof of Proposition 5. We start by considering the case $d = 2$ and then study the more general neural space $\mathcal{N}_\sigma(C^{1,h}, C^{h\times k})$.

Recall $\lambda(C) = \operatorname{Span}\{x \mapsto id_{\mathbb{R}^X} \cdot x\}$. Elements in $C^{1,h}$ can be represented as affine maps $x \mapsto Bx + c$ where B and c have the following block representations

$$B = \begin{bmatrix} b_1 \operatorname{id} \\ \vdots \\ b_h \operatorname{id} \end{bmatrix} \quad \text{and} \quad c = \begin{bmatrix} c_1 \mathbb{1} \\ \vdots \\ c_h \mathbb{1} \end{bmatrix}.$$

While elements in $C^{h,k}$ can be represented as affine maps $x \mapsto Ax + d$ where $d \in \mathbb{R}$ and

$$A = \begin{bmatrix} a_{1,1} \cdot \operatorname{id}_{\mathbb{R}^X} & \cdots & a_{1,h} \cdot \operatorname{id}_{\mathbb{R}^X} \\ \vdots & \vdots & \vdots \\ a_{k,1} \cdot \operatorname{id}_{\mathbb{R}^X} & \cdots & a_{h,h} \cdot \operatorname{id}_{\mathbb{R}^X} \end{bmatrix} = \tilde{A} \otimes \operatorname{id}_{\mathbb{R}^X},$$

where $\tilde{A} = [a_{i,j}] \in \mathbb{R}^{k\times h}$.

Given $i \in X$ and $s = 1, \ldots, h$, we can write elements $\theta \in \mathcal{N}_\sigma(C^{1,h}, C^{h,k})$ as

$$\theta_{s,i}(x) = A\sigma(Bx + c) = \sum_{j=1}^{h} a_{s,j}\sigma\left(b_j x_i + c_j\right)$$

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

$$\theta_{s,i}(x) = \sum_{j=1}^{h} a_{s,j}\sigma\left(b_j x_i + c_j\right) = \xi_s(x) \tag{12}$$

where

$$\xi_s(y) := \sum_{j=1}^{h} a_{s,j}\sigma\left(b_j y + c_j\right).$$

That is, $\xi \in \mathcal{N}_\sigma(\mathbb{R}^m, \mathbb{R}^h, \mathbb{R}^k)$. In other words, taking the limit as $h \to \infty$ and setting $k = 1$, we obtain the proof of the theorem for the case $d = 2$. For $d > 2$, it suffices to note that the composition of spaces of the type $\mathcal{N}_\sigma(C^{1,h}, C^{h\times k})$ again yields elements of the same type. This concludes the proof. □

C.2 Entry-wise Separation Constrained Universality

Let $M_i < \mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X)$ for $i = 1, \dots, d$ and $P_x := \pi_{x*} M_{d+1}$, we obtain

$$\pi_{x*}\,\mathcal{U}_\sigma(M_1, \dots, M_d, M_{d+1}) = \mathcal{U}_\sigma(M_1, \dots, M_d, \pi_{x*} M_{d+1}) = \mathcal{U}_\sigma(M_1, \dots, M_d, P_x).$$

Note that $P_x < \mathrm{Aff}_{G_x}(\mathbb{R}^X, \mathbb{R})$ thanks to Remark 3. Let $X_1, \dots, X_\ell$ be the orbits of G_x over X. Then

$$\mathrm{Hom}_{G_x}(\mathbb{R}^X, \mathbb{R}) = \mathrm{Hom}_{G_x}(\mathbb{R}^{X_1 \sqcup \cdots \sqcup X_\ell}, \mathbb{R}) = \mathrm{Hom}_{G_x}(\mathbb{R}^{X_1} \oplus \cdots \oplus \mathbb{R}^{X_\ell}, \mathbb{R}) = \mathrm{Hom}_{G_x}(\mathbb{R}^{X_1}, \mathbb{R}) \oplus \cdots \oplus \mathrm{Hom}_{G_x}(\mathbb{R}^{X_\ell}, \mathbb{R}).$$

Then there are projections

$$\alpha_i : \mathrm{Hom}_{G_x}(\mathbb{R}^X, \mathbb{R}) \to \mathrm{Hom}_{G_x}(\mathbb{R}^{X_i}, \mathbb{R}), \quad i = 1, \dots, \ell.$$

We define

$$I_i < \mathrm{Aff}_{G_x}(\mathbb{R}^{X_i}, \mathbb{R}) < \mathrm{Aff}_{G_x}(\mathbb{R}^X, \mathbb{R}),$$

where the linear component of I_i is given by $\alpha_i \lambda(P_x)$, and its translation component equals $\tau(P_x)$. In particular, since G_x is the stabilizer of x in X, one of the orbits is the singleton $\{x\}$ itself. Without loss of generality, we assume $X_1 = \{x\}$. We also assume that $\mathrm{id} \in M$, in which case $I_1 = L$, where $L = \mathrm{Aff}(\mathbb{R}, \mathbb{R})$.

We now prove Theorem 2. For this purpose, we will make use of the following lemmas.

Lemma 4. *With the notation of Theorem 2, and setting for simplicity $P := P_{x_i}$ and $\rho := \rho_i$, for any $d \geq 1$ the following separation-constrained universality holds*

$$\mathcal{U}_\sigma(M_1, ..., M_d, P, I_1) = \mathcal{C}_\rho(V). \tag{13}$$

Proof of Lemma 4. Recall that every $f \in \mathrm{Aff}(V, W)$ can be uniquely decomposed as $f = \tau_w \circ \phi$ with $\tau_w(w') = w' + w$, for some $\phi \in \mathrm{Hom}(V, W)$ and $w, w' \in W$. From Pacini et al. (2024a), we have that the map f is G-equivariant precisely when ϕ is G-equivariant and v belongs to the fixed-point subspace $W^G = \{\, w \in W \mid gw = w \text{ for all } g \in G \,\}$. In particular, there is a natural linear projection $\lambda : \mathrm{Aff}_G(V, W) \to \mathrm{Hom}_G(V, W)$ which associates to an equivariant affine map its linear component.

Write

$$\lambda(P) = \mathrm{Span}\{\phi^1, \dots, \phi^\ell\} \subseteq \mathrm{Hom}_{G_x}(\mathbb{R}^X, \mathbb{R}).$$

Note that $I_1 = L$ (Example 1.i). Elements in $\mathcal{N}_\sigma(P^{1\times k}, L^{k\times 1})$ can be represented as maps

$$\eta : v \mapsto a^\top \sigma(Bv + c)$$

where a, B and c have the following block representations

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix}, \quad B = \begin{bmatrix} b_{1,1}\phi^1 + \cdots + b_{1,m}\phi^m \\ \vdots \\ b_{k,1}\phi^1 + \cdots + b_{k,m}\phi^m \end{bmatrix}, \quad \text{and} \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix}.$$

Therefore, we can write

$$\eta(v) = \sum_{i=1}^{k} a_i \sigma \left(\sum_{j=1}^{\ell} b_{i,j} \phi^j(v) + c_i \right) = \zeta(\phi^1(v), \dots, \phi^\ell(v))$$

where $\zeta \in \mathcal{N}_\sigma(L^{\ell\times k}, L^{k\times 1})$. In a similar way, any function $\eta \in \mathcal{N}_\sigma(P^{h\times k}, L^{k\times 1})$ can be written as

$$\eta(v) = \zeta(\phi^{1,1}(v), \dots, \phi^{\ell,1}(v), \dots, \phi^{1,h}(v), \dots, \phi^{\ell,h}(v)) \tag{14}$$

where $\lambda(P^{h\times 1}) = \mathrm{Span}\{\phi^{1,i}, \dots, \phi^{\ell,i}\}_{i=1,\dots,h}$. Any function

$$\xi \in \mathcal{N}_\sigma(M_1^{1\times h_1}, \dots, M_d^{h_d\times h}, P^{h\times k}, L^{k\times 1})$$

can be written as

$$\xi = \eta \circ \sigma \circ \theta$$

where $\eta \in \mathcal{N}_\sigma(P^{h\times k}, L^{k\times 1})$ and $\theta \in \mathcal{N}_\sigma(M_1^{1\times h_1}, \dots, M_d^{h_d\times h})$. Plugging this into Equation 14, we obtain

$$\eta \circ \sigma \circ \theta(w) = \zeta\left[\left(\phi^{1,i}(\sigma \circ \theta(w)), \dots, \phi^{\ell,i}(\sigma \circ \theta(w)\right)_{i=1,\dots,h}\right].$$

Note that

$$\rho = \rho\left(\left\{\phi^{j,i} \circ \sigma \circ \theta \mid \theta \in \mathcal{N}_\sigma(M_1^{1\times h_1}, \dots, M_d^{h_d\times h}), i \in [h], j \in [\ell]\right\}\right)$$

since $\phi^{1,i}, \dots, \phi^{\ell,i}$ is a basis for $\lambda(P^{h\times 1})$ for each $i = 1, \dots, h$. We conclude the proof by following the argument of Theorem 1, applying Lemma 3 to θ with rational parameters, and observing that $\zeta \in \mathcal{U}_\sigma(L^{\ell h\times 1}, L) = \mathcal{C}(\mathbb{R}^{\ell h})$. □

For brevity, we will state all subsequent propositions and lemmas in terms of $\mathcal{U}_\sigma(\underbrace{M, ..., M}_{d \text{ times}}, P)$ or similar forms. The same statements, however, extend verbatim to the more general setting $\mathcal{U}_\sigma(M_1, \dots, M_d, \underbrace{M, ..., M}_{d \text{ times}}, P)$.

Proof of Theorem 2. Thanks to Equation 10, it suffices to prove that

$$\mathcal{U}_\sigma(\underbrace{M, ..., M}_{d+1 \text{ times}}, P_{x_i}) = \mathcal{C}_{G_{x_i},\rho_i}(\mathbb{R}^X, \mathbb{R}), \quad i = 1, \dots, s.$$

In the following, we drop the indices for simplicity of exposition. First note that

$$\mathcal{U}_\sigma(\underbrace{M, ..., M}_{d+1 \text{ times}}, I_1) \subseteq \mathcal{U}_\sigma(\underbrace{M, ..., M}_{d+1 \text{ times}}, P) \subseteq \mathcal{C}_\rho(V), \tag{15}$$

where the first inclusion second inclusion follows because $I_1 \subseteq P$ and the last inclusion comes from the definition of ρ. For the inclusions in the opposite direction, note that

$$\mathcal{C}_\rho(V) = \mathcal{U}_\sigma(\underbrace{M, ..., M}_{d \text{ times}}, P, I_1) \subseteq \mathcal{U}_\sigma(\underbrace{M, ..., M}_{d \text{ times}}, M, I_1), \tag{16}$$

where the first equality is true by Lemma 4 and the following inclusion holds because M is more expressive than P by definition. Now using Equations 15 and 16 the claim directly follows. □

Proof of Theorem 3. Note that

$$\pi_{1^*}\mathcal{U}_\sigma(M_1, \dots, M_f, C) = \mathcal{U}_\sigma(M_1, \dots, M_f, I_1) = \mathcal{U}_\sigma(M_1, \dots, M_{f-1}, \pi_1 M_f, I_1).$$

The proof follows directly from Lemma 4. □

Lemma 5. *The projection $\pi_i^*\,\mathcal{N}_\sigma(C, P, C)$ separates $\mathrm{Stab}_{S_n}(i)$-orbits in $\mathbb{R}^n$.*

Proof. Note that in general

$$\mathcal{N}_\sigma(M_1, \dots, M_d) = \mathcal{N}_\sigma(M_1, \dots, M_{i-1}, C)\,\hat{\circ}\,\mathcal{N}_\sigma(M_i, \dots, M_d)$$

We start by addressing the general invariant case first, namely:

$$\mathcal{N}_\sigma(M^{1\times h}, N^{h\times k}, I^{k\times 1}) = \mathcal{N}_\sigma(M^{1\times h}, C^{h\times k})\,\hat{\circ}\,\mathcal{N}_\sigma(N^{h\times k}, I^{k\times 1}), \tag{17}$$

Thanks to Equation 17 we can focus on writing down distinct formulas for functions in both $\mathcal{N}_\sigma(N^{h\times k}, I^{k\times 1})$ and $\mathcal{N}_\sigma(M^{1\times h}, C^{h\times h})$.

Consider the space $\mathcal{N}_\sigma(N^{h\times k}, I^{k\times 1})$.

Let $n = \dim \lambda(N)$ and let $\phi^1, \dots, \phi^{hn}$ be a basis for $N^{h\times 1}$ and $\lambda(I) = \mathrm{Span}\{x \mapsto \mathbb{1}^t \cdot x\}$. Elements in $N^{h\times k}$ 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 $I^{h\times 1}$ can be represented as affine maps $x \mapsto Ax + d$ where $d \in \mathbb{R}$ and

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

Denote by ϕ_i^j the projection of the i-th component of the function ϕ^j. We can write elements $\eta \in \mathcal{N}_\sigma(N^{h\times k}, I^{k\times 1})$ as

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

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

$$\eta(x) = \sum_{j=1}^{k} a_j \sum_{i\in Y} \sigma\left(\sum_{t=1}^{hn} b_{j,t}\phi_i^t(x) + c_j\right) = \tag{18}$$

$$\sum_{i\in Y}\sum_{j=1}^{k} a_j\sigma\left(\sum_{t=1}^{hn} b_{j,t}\phi_i^t(x) + c_j\right) = \sum_{i\in Y} \zeta(\phi_i^1(x), \dots, \phi_i^{hn}(x)) \tag{19}$$

where

$$\zeta(y_1, \dots, y_{hn}) := \sum_{j=1}^{k} a_j\sigma\left(\sum_{t=1}^{hn} b_{j,t}y_t + c_j\right)$$

is a standard multilayer perceptron in $\mathcal{N}_\sigma(L^{hn\times k}, L^{k\times 1})$.

Consider now the space $\mathcal{N}_\sigma(M^{1\times h}, C^{h\times h})$.

Let $m = \dim \lambda(M)$ and let $\psi^1, \dots, \psi^m$ be a basis for M and $\lambda(C) = \mathrm{Span}\{x \mapsto id_{\mathbb{R}^X} \cdot x\}$. Elements in $M^{1\times h}$ can be represented as affine maps $x \mapsto Bx + c$ where B and c have the following block representations

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

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

$$A = \begin{bmatrix} a_{1,1}\cdot id_{\mathbb{R}^X} & \cdots & a_{1,h}\cdot id_{\mathbb{R}^X} \\ \vdots & \vdots & \vdots \\ a_{h,1}\cdot id_{\mathbb{R}^X} & \cdots & a_{h,h}\cdot id_{\mathbb{R}^X} \end{bmatrix} = \tilde{A} \otimes id_{\mathbb{R}^X},$$

where $\tilde{A} = [a_{i,j}]_{i,j=1,\dots,h}$.

Denote by ψ_i^j the projection of the i-th component of the function ψ^j. Given $i \in X$ and $s = 1, \dots, h$, we can write elements $\theta \in \mathcal{N}_\sigma(M^{1\times h}, C^{h\times h})$ as

$$\theta_{s,i}(x) = A\sigma(Bx + c) = \sum_{j=1}^{h} a_{s,j}\sigma\left(\sum_{t=1}^{m} b_{j,t}\psi_i^t(x) + c_j\right)$$

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

$$\theta_{s,i}(x) = \sum_{j=1}^{h} a_{s,j}\sigma\left(\sum_{t=1}^{m} b_{j,t}\psi_i^t(x) + c_j\right) = \xi_s(\psi_i^1(x), \ldots, \psi_i^m(x)) \tag{20}$$

where

$$\xi_s(y_1, \ldots, y_m) := \sum_{j=1}^{h} a_{s,j}\sigma\left(\sum_{t=1}^{m} b_{j,t}y_t + c_j\right).$$

Namely, $\xi \in \mathcal{N}_\sigma(L^{m\times h}, L^{h\times s})$.

Consider now the composition $\mathcal{N}_\sigma(M^{1\times h}, N^{h\times k}, I^{k\times 1})$.

Each element in $\mathcal{N}_\sigma(M^{1\times h}, N^{h\times k}, I^{k\times 1})$ is written as

$$\eta \circ \theta(x) = \sum_{i\in Y} \zeta(\phi_i^{1,1}(\theta_{1,*}(x)), \ldots, \phi_i^{h,n}(\theta_{1,*}(x))) = \tag{21}$$

$$\sum_{i\in Y} \zeta(\phi_i^{1,1}([\xi_1(\psi_j^1(x), \ldots, \psi_j^{hm}(x))]_{j\in X}), \ldots, \phi_i^{h,n}([\xi_h(\psi_j^1(x), \ldots, \psi_j^{hm}(x))]_{j\in X})). \tag{22}$$

Consider the case

$$M = \mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}^n) = \left\{v \mapsto (\lambda I + \mu \mathbb{1}^\top \mathbb{1})v + y\mathbb{1}_{[n]} \;\middle|\; \lambda, y \in \mathbb{R}\right\}.$$

In this case we have a basis defined by

$$\phi^1 = I \text{ and } \phi^2 = \mathbb{1}^\top \mathbb{1}.$$

In particular, for each $i = 1, \ldots, n$:

$$\phi_i^1 = e_i^\top \text{ and } \phi_i^2 = \mathbb{1}^\top.$$

Therefore, for $x \in \mathbb{R}^n$:

$$\tilde{\phi}_i(x) = (e_i^\top \cdot x, \mathbb{1}^\top \cdot x),$$

or, alternatively, for $x = (x_1, \ldots, x_n)$,

$$\tilde{\phi}_i(x) = (x_i, x_1 + \cdots + x_n).$$

- We want to write elements in $\mathcal{N}_\sigma(M^{1\times h}, I^{h\times 1})$, by Equation 18, we have

  $$\eta(x) = \sum_{i=1}^{n} \zeta(x_i, x_1 + \cdots + x_n),$$

  for some neural network $\zeta \in \mathcal{N}_\sigma(L^{2\times h}, L^{h\times 1})$. Similarly, in the case $\mathcal{N}_\sigma(M^{k\times h}, I^{h\times 1})$, for $x^1, \ldots, x^k \in \mathbb{R}^n$, we can write

  $$\eta(x_1, \ldots, x_k) = \sum_{i=1}^{n} \zeta(x_{1,i}, \ldots, x_{k,i}, \overline{x}_1, \ldots, \overline{x}_k),$$

  where $\overline{x}_i = x_{i,1} + \cdots + x_{i,n}$ for $i = 1, \ldots, k$ and $\zeta \in \mathcal{N}_\sigma(L^{2k\times h}, L^{h\times 1})$.

- Now, we want to write functions in $\mathcal{N}_\sigma(M^{1\times k}, C^{k\times k})$. Thanks to Equation 12, we can write for $s = 1, \ldots, k$ and $i = 1, \ldots, n$:

  $$\theta_{s,i}(x) = \xi_s(x_i, x_1 + \cdots + x_n)$$

  for $\xi \in \mathcal{N}_\sigma(L^{2\times k}, L^{k\times k})$.

- Compute $\mathcal{N}_\sigma(M^{1\times k}, M^{k\times h}, I^{h\times 1}) = \mathcal{N}_\sigma(M^{1\times k}, C^{k\times k}) \,\hat{\circ}\, \mathcal{N}_\sigma(M^{k\times h}, I^{h\times 1})$ by composition:

  $$\eta \circ \theta(x) = \sum_{i=1}^{n} \zeta(\xi_1(x_i, \overline{x}), \ldots, \xi_k(x_i, \overline{x}), \overline{\xi}_1(x), \ldots, \overline{\xi}_k(x))$$

  where we write $\overline{x} = x_1 + \cdots + x_n$ and $\overline{\xi}_i(x) = \xi_i(x_1, \overline{x}) + \cdots + \xi_i(x_n, \overline{x})$ for $i = 1, \ldots, k$.

  Note that:

- For $k \to \infty$, we choose ξ_j to approximate the polynomial $(x, y) \mapsto x^j$ for $j = 1, \ldots, k$.
- Moreover, for $h \to \infty$ and $k \geq n$, we can choose ζ to approximate any function $(x_1, \ldots, x_k, y_1, \ldots, y_k) \mapsto \tilde{\zeta}(y_1, \ldots, y_n)$ for any continuous function $\tilde{\zeta} \in \mathcal{C}(\mathbb{R}^n)$.

Therefore,

$$\eta \circ \theta(x) = n \cdot \tilde{\zeta}(x_1 + \cdots + x_n, \ldots, x_1^n + \cdots + x_n^n).$$

We know that these functions are permutation invariant universal.

- Define $P = \pi_1^*(\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R}^n))$, where π_1^* is the pushforward map defined in Proposition 4 for $x = 1 \in [n]$.

  Let $\overline{e}_1 := e_2 + \cdots + e_n$. Note that

  $$P = \mathrm{Aff}_{S_{(n-1,1)}}(\mathbb{R}^n, \mathbb{R}) = \left\{ x \mapsto (\lambda e_1 + \mu \overline{e}_1)^\top \cdot x + c \mid \lambda, \mu, c \in \mathbb{R} \right\},$$

  then

  $$A = \begin{bmatrix} a_{1,1}e_1 + a_{1,2}\overline{e}_1 \\ \vdots \\ a_{h,1}e_1 + a_{h,2}\overline{e}_1 \end{bmatrix}^\top.$$

  Note that

  $$\eta(x) = \tag{23}$$

  $$\sum_{j=1}^{h} \left[a_{j,1}\sigma\left(\sum_{t=1}^{m} b_{j,t}\phi_1^t(x) + c_j \right) + \sum_{i=2}^{n} a_{j,2}\sigma\left(\sum_{t=1}^{m} b_{j,t}\phi_i^t(x) + c_j \right) \right] = \tag{24}$$

  $$\zeta_1(\phi_1^1(x), \ldots, \phi_1^m(x)) + \sum_{i=2}^{n} \zeta_2(\phi_i^1(x), \ldots, \phi_i^m(x)), \tag{25}$$

  where

  $$\zeta_s(y_1, \ldots, y_n) := \sum_{j=1}^{h} a_{j,s}\sigma\left(\sum_{t=1}^{h} b_{j,t}y_t + c_j \right)$$

  for $s = 1, 2$. We want to write elements in $\mathcal{N}_\sigma(M^{1\times h}, P^{h\times 1})$, by adapting Equation 23, we have

  $$\eta(x) = \zeta_1(x_1, x_1 + \cdots + x_n) + \sum_{i=2}^{n} \zeta_2(x_i, x_1 + \cdots + x_n)$$

  for some neural network $\zeta_1, \zeta_2 \in \mathcal{N}_\sigma(L^{2\times h}, L^{h\times 1})$.

- We want to write elements in $\mathcal{N}_\sigma(M^{k\times h}, P^{h\times 1})$:

  $$\eta(x) = \zeta_1(x_{1,1}, \ldots, x_{k,1}, \overline{x}_1, \ldots, \overline{x}_k) + \sum_{i=2}^{n} \zeta_2(x_{1,i}, \ldots, x_{k,i}, \overline{x}_1, \ldots, \overline{x}_k)$$

  for some neural network $\zeta_1, \zeta_2 \in \mathcal{N}_\sigma(L^{2\times h}, L^{h\times 1})$.

- Compute $\mathcal{N}_\sigma(C^{1\times k}, M^{k\times h}, P^{h\times 1}) = \mathcal{N}_\sigma(C^{1\times k}, C^{k\times k}) \,\hat{\circ}\, \mathcal{N}_\sigma(M^{k\times h}, P^{h\times 1})$ by composition. With similar computations as above we obtain:

  $$\eta(x) =$$
  $$\zeta_1(\xi_1(x_1), \ldots, \xi_k(x_1), \overline{\xi}_1(x), \ldots, \overline{\xi}_k(x)) +$$
  $$\sum_{i=2}^{n} \zeta_2(\xi_1(x_i), \ldots, \xi_k(x_i), \overline{\xi}_1(x), \ldots, \overline{\xi}_k(x))$$

  where $\overline{\xi}_j(x) = \sum_{i=1}^{n} \xi_j(x_i, \overline{x})$ for each $j = 1, \ldots, k$.

- For $k \to \infty$, we can approximate
$$\xi_j(x) \to [x \mapsto x^j]$$
for $j = 1, \ldots, k$.
- Choose ζ_1 to approximate any continuous function and set ζ_2 everywhere zero, meaning we can choose the readout output M to be C.

Then we obtain that

$$\eta(x) = \zeta_1(x_1, x_1 + \cdots + x_n, \ldots, x_1^n + \cdots + x_n^n).$$

This set is universal in $\mathcal{C}_{S_{(n-1,1)}}(\mathbb{R}^n)$ where $S_{(n-1,1)} < S_n$ is the stabilizer of 1 inside $[n]$.

□