
UNIVERSITÀ
DI TRENTO

*Department of Information Engineering and Computer Science*

ICT International Doctoral School

Cycle XXXVII

Academic Year 2024/2025

---

# Theoretical Properties of Equivariant Neural Networks

---

**PhD Candidate:** Marco Pacini

**Advisor:** Bruno Lepri

# Contents

# Chapter 1

# Introduction

Modern neural networks can often interpolate the training data while still generalizing well [74, 10]. This strong performance in the interpolation regime motivates the study of the expressivity of deep learning models, and encourages the search for architectures with greater expressive power. A natural viewpoint for studying expressivity is approximation theory, which asks how well a given function class can approximate a prescribed target family. In particular, it is important to understand how design choices and hyperparameters, such as depth, intermediate representations, and symmetry constraints, affect expressivity. The main goal of this thesis is to develop a mathematical account of these architectural effects on approximation. We focus on equivariant architectures, where modeling assumptions can be stated precisely in representation-theoretic terms. This provides a principled framework to compare the expressivity of architectures and to derive design rules by linking architectural choices to the approximation properties of the resulting hypothesis spaces.

## 1.1 Statistical Learning

The above considerations can be clarified and formalized within a standard statistical learning framework for excess risk minimization, which we briefly review next; see [111] and [19] for further details. In particular, in this thesis we focus on supervised learning, in which we observe labeled data and aim to learn a map from inputs to outputs. Concretely, we consider an input space $\mathcal{X}$, an output space $\mathcal{Y}$, and a data-generating distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. A predictor is a measurable map $f : \mathcal{X} \to \mathcal{Y}$, and we evaluate predictions through a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, where $\ell(\hat{y}, y)$ quantifies the cost of predicting $\hat{y}$ when the true label is $y$. The *population risk* of $f$ is then

$$\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f(x), y)\big],$$

and statistical learning can be viewed as the problem of finding a predictor with small population risk. In practice, $\mathcal{D}$, and hence $\mathcal{R}(f)$, are unknown.

Accordingly, $\mathcal{R}(f)$ is typically estimated using the *empirical risk*

$$\widehat{\mathcal{R}}_S(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i),$$

computed from a finite dataset $S := \{(x_i, y_i)\}_{i=1}^n$ of i.i.d. samples from $\mathcal{D}$. A learning algorithm typically restricts attention to a hypothesis space

$$\mathcal{H} \subseteq \{f : \mathcal{X} \to \mathcal{Y}\}$$

and seeks a predictor that minimizes, or approximately minimizes, empirical risk within $\mathcal{H}$.

To connect these notions to an optimal target, let $f^\star$ denote a minimizer of the population risk, known as a Bayes predictor. That is, $f^\star$ is any measurable map satisfying

$$f^\star \in \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}(f).^1$$

Since $f^\star$ may not belong to $\mathcal{H}$, we introduce the *best-in-class population minimizer*

$$\bar{f} \in \arg\min_{f\in\mathcal{H}} \mathcal{R}(f).$$

In the language of approximation theory, $\bar{f}$ is the best approximation of $f^\star$ available in $\mathcal{H}$. However, as noted above, the population risk $\mathcal{R}(f)$ is unknown and therefore cannot be minimized directly. In many classical machine learning scenarios, empirical risk minimization leads to a convex problem that is amenable to exact optimization. By contrast, neural network training is typically non-convex, and is therefore carried out using gradient-based methods that only approximately minimize the empirical risk. To account for this, we let $\tilde{f}$ denote the output of a specific training algorithm $\mathcal{A}$ given the hypothesis space $\mathcal{H}$ and the dataset $S$. This is the predictor actually produced in practice. A classical narrative decomposes the resulting excess risk as

$$\begin{aligned}
\mathcal{R}(\tilde{f}) - \mathcal{R}(f^\star) &= && \textbf{(Excess risk)} \\
&= \mathcal{R}(\bar{f}) - \mathcal{R}(f^\star) && \textbf{(Approximation error)} \\
&+ \mathcal{R}(\tilde{f}) - \mathcal{R}(\bar{f}). && \textbf{(Learning error)}
\end{aligned}$$

Here, the approximation error $\mathcal{R}(\bar{f}) - \mathcal{R}(f^\star)$ quantifies the mismatch between the target and the hypothesis space. On the other hand, the *learning error* captures the effect of finite samples together with the selection algorithm.[2]

In this thesis we focus on the first term, the approximation error. In particular, we study *universality*, namely the ability of a model class, under prescribed inductive biases, to approximate any continuous target function with arbitrary

[1] For the sake of simplicity, we assume from now on that all minimizers exist.

[2] This is not standard terminology; related quantities are often discussed under the names *estimation error* and, more broadly, algorithm-dependent generalization terms. More details in Chapter 8.

precision on compact sets in the uniform norm. This makes clear that a universal model can drive the approximation error to zero, provided the target function $f^\star$, which we assume to be a Bayes predictor, is continuous and compatible with the prescribed inductive biases.

Universality is a *sufficient* condition for achieving arbitrarily small approximation error, which in turn is a *necessary* condition for learning with arbitrary precision.

In particular, universality provides a clean and widely used notion of expressivity, and it is often invoked to reason about the approximation component of the excess risk. However, the statistical learning framework also highlights several limitations of universality as a guiding principle. We highlight three aspects that would motivate more adequate notions of expressivity and its effects.

- Universality is typically an asymptotic property, quantifying what a model class can achieve in the limit of growing capacity. In deep learning practice, models have finite width and depth, and without approximation-rate guarantees there is no reason to expect the approximation error to be small. As a consequence, universality may hold while still being uninformative about whether a given architecture can approximate the relevant targets at the scale imposed by the task and computational budget.
- Uniform approximation controls the worst-case error over the entire input domain. In supervised learning, however, performance is typically evaluated through a risk $\mathcal{R}(f)$, which weights errors according to the data distribution on $\mathcal{X}$ and the noise in $\mathcal{Y}$. When observations are noisy or when the input distribution concentrates on a low-measure subset of $\mathcal{X}$, requiring small error uniformly on a large compact set may be unnecessarily demanding and may not reflect the relevant notion of closeness for learning. Moreover, the population risk $\mathcal{R}(f)$ depends on the choice of loss function $\ell$. When $\ell$ is not induced by the uniform norm, as is the case for the cross-entropy and mean squared error losses, proximity in the uniform sense need not align with proximity as measured by the loss.
- Universality asks to approximate an entire function space, for instance all continuous functions satisfying a prescribed inductive bias. Yet, in supervised learning the targets of interest are typically Bayes predictors, or functions that are close to them. Therefore, an alternative notion of universality could instead focus on approximating only these target families, rather than the full class of continuous functions compatible with the bias.

Finally, it is not generally understood how achieving universality, or more broadly improving approximation error, affects the estimation and optimization errors. However, contrary to the classical intuition that more expressive models should overfit, highly expressive neural networks often generalize well in practice, further motivating the study of expressivity, since it appears to also alleviate estimation and optimization errors.

## 1.2 Inductive Biases in Machine Learning

A recurring lesson in deep learning practice is that performance improvements rarely come from optimization alone; rather, they arise from guiding the learning pipeline toward inductive biases that reflect the task structure known a priori. This has the effect of reducing the estimation error and, thus, improving sample efficiency. However, inductive biases appear in several forms. Natural sciences such as physics, chemistry, and biology often describe phenomena via dynamical systems and differential equations [67, 96, 31]; this prior structure can be leveraged as an inductive bias in machine learning models. This has sparked a broad research area known as physics-informed machine learning [59]. As a result, this line of work has found applications ranging from turbulence modeling [72] to materials design [92] and epidemic forecasting [97].

A second class of inductive biases encompass relational structures [5], leveraging the fact that data are naturally organized as entities and relations, as in graph learning [104, 112, 65, 33] and neuro-symbolic computing [30, 110, 18, 75]. Lastly, the class of inductive biases of main interest in this thesis arises in symmetry-preserving tasks. For example, such structures are exploited by classical convolutional networks for translation symmetry [70], as well as equivariant architectures for sets [95, 124], graphs [78], and geometric data [57, 121, 24]. Such inductive biases are typically introduced via data augmentation, additional penalty terms in the loss, or architectural design. Among these, the latter is the only mechanism that directly restricts the hypothesis space, which makes it a natural choice when one aims to enforce the assumed structure and improve robustness and reliability. For this reason, this thesis focuses on architectural inductive biases that encode symmetry. In particular, the next section discusses how to encode symmetry as an architectural bias, and how this can introduce undesired artifacts, which will be the leitmotif throughout this manuscript.

## 1.3 Symmetry as Inductive Bias

Equivariance formalizes the idea that transforming the input should produce a predictable transformation of the output, and it provides a principled inductive bias by restricting the hypothesis space to functions compatible with the symmetries preserved by the task. The first efforts to develop machine learning models that preserve symmetries predate modern deep learning. Early connections between equivariance and machine learning can be traced back at least to the work of [57]. Within deep learning, a first systematic effort to formulate equivariant neural networks and to analyze their structure already appears in [121]. Only in recent years, with the rise of deep learning [69] as a main tool to learn from data at scale in weakly structured and noisy tasks, has the need for symmetry-preserving neural networks led to a large body of work on equivariant architectures [25, 64], now commonly placed under the umbrella name of geometric deep learning [21].

Such ideas have been adapted to a range of data modalities, including point

clouds [38], graphs [115], differential manifolds [60], and simplicial complexes [6]. As a result, equivariant models have studied across several symmetry-sensitive domains, from high-energy physics [17] and structural biology or drug discovery [55] to robotics [53, 50, 49], calibration [13], reinforcement learning [117], and medical imaging [66]. To construct these models, some general-purpose approaches rely on symmetrization operators which can cause the dimension of intermediate representations to grow rapidly making them impractical for large groups or high-resolution inputs [98]. Canonicalization can mitigate this cost in some settings [93, 56, 113], but is not always applicable as shown by [29]. Other approaches trade exactness for efficiency by enforcing approximate equivariance [35, 89]. However, the dominant paradigm in geometric deep learning is to build equivariant models by composing building blocks that are equivariant by themselves, since equivariance is preserved under composition. This includes architectures based on group convolutions and steerable layers [25, 64, 68] or polynomial and tensor features [16, 114, 63, 119, 44, 116].

Despite its simplicity and empirical success, this construction paradigm can introduce unintended artifacts and unexpected behaviors. More precisely, equivariance interacts non-trivially with standard design choices in deep learning, namely the use of linear layers and pointwise nonlinearities. This standard approach, equivariant linear maps composed with pointwise activations, is convenient and underlies many practical architectures [25, 21]. However, for a given choice of representations, pointwise nonlinearities may fail to define non-trivial equivariant nonlinear maps, thereby heavily constraining the design space. This issue was already noted in early work that explicitly classified equivariant networks under pointwise activations [121]. At the same time, several practically relevant nonlinear features are not point-wise, including norm nonlinearities, squashing nonlinearities, tensor-product nonlinearities, and gated nonlinearities [122, 100, 63, 119]. Despite the popularity of these methods, frameworks to study the properties of such nonlinearities are still largely absent [118].

The work presented in this thesis is dedicated to understanding these artifacts and their consequences. First, we study the interaction between pointwise activations and equivariance; in particular, we characterize which pairs of representations and pointwise activations induce non-trivial equivariant nonlinearities, thereby identifying when the resulting hypothesis spaces are non-degenerate. Second, we study expressivity limitations that may arises in equivariant neural networks. Indeed, despite the appeal of equivariant models, their expressivity is often limited by how well they can distinguish non-equivalent inputs, a property known as separation power. In permutation-equivariant settings, separation is commonly analyzed through the Weisfeiler–Leman (WL) test, which provides a fundamental proxy for the separation power of graph neural networks [105, 82, 7, 15] and invariant graph networks [42, 77, 43]. More recently, homomorphism counting has been studied as a more fine-grained approach to separation [125, 2, 37, 14]. Beyond graph domains, [54] extends WL test to to geometric graphs. In this thesis, we move beyond the graph setting to a general equivariant framework in which WL-based approaches are inapplicable. In particular, we study networks built from regular $G$-convolutions [25], a

class that includes several widely used models such as IGNs [78], circular CNNs [98], and icosahedral CNNs [26].

## 1.4 Contributions

The main contributions of this thesis are as follows.

- We characterize which pairs of pointwise activations and representations induce equivariant nonlinearities, building on [83].
- We study the separation power of families of neural networks with a prescribed architecture. In particular, we describe how changes in architectural structure and hyperparameters may or may not affect separation power, see [84].
- We study the role of depth in equivariant architectures, and show that without sufficient depth or an appropriate readout, these families cannot approximate all continuous functions compatible with the prescribed separation constraint, see [86].
- We study universality for equivariant neural networks under a prescribed separation constraint, and show that this form of universality can always be achieved given sufficient depth or suitable readout layers, building on [85].

## 1.5 Thesis Structure

We conclude this introduction with a brief road-map of the thesis. Chapter 2 introduces the basic material needed throughout the thesis, including group theory and a brief introduction to representation theory. We then turn to a key architectural ingredient: Chapter 3 develops the theory of real activations, characterizing which activations induce equivariant nonlinearities. With these tools in place, Chapter 4 introduces the notation and structures used in the rest of the thesis, including neural spaces and universality classes, which provide a formal language to encode neural network architectures and their associated hypothesis spaces. We then study expressivity from an approximation-theoretic viewpoint. Indeed, Chapter 5 develops instruments to assess the separation power of neural spaces, a prerequisite for analyzing approximation capabilities. Building on this, Chapter 6 characterizes universality classes for shallow invariant architectures and shows that separation alone does not always suffice to describe the functions that can be approximated under symmetry constraints. Finally, Chapter 7 shows how this limitation can be removed by increasing depth, leading to universality for all continuous functions compatible with the prescribed separation constraint. The thesis concludes with Chapter 8, which summarizes the main findings and discusses limitations and future directions. Additional technical material and proofs are deferred to the appendices.

# Chapter 2

# Preliminaries

We start by introducing the necessary preliminaries in group theory and representation theory, which will be used throughout the rest of the manuscript.

## 2.1 Groups and Homomorphisms

**Definition 2.1.1.** A *group* is a pair $(G, \cdot)$ where $G$ is a set and $\cdot : G \times G \to G$ is a function satisfying the following axioms.

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

A group is *finite* if it contains a finite number of elements. A group is *abelian* or *commutative* if $gh = hg$ for each $g, h \in G$. If a group $H$ is contained in a group $G$, then $H$ is called a *subgroup* of $G$, and we write $H < G$.

**Example 2.1.2.** Here we present some fundamental examples of groups.

- The sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$ are groups under addition and form a chain of subgroups $\mathbb{Z} < \mathbb{Q} < \mathbb{R} < \mathbb{C} < \mathbb{H}$.
- The sets $\mathbb{Q}^*$, $\mathbb{R}^*$, $\mathbb{C}^*$, and $\mathbb{H}^*$ of invertible elements in $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$, respectively, are groups under multiplication and form a chain of subgroups as well. Note that $\mathbb{Z}^*$ is not a group under multiplication, since its only invertible elements are $\pm 1$.
- Let $X$ be a set and define the set of permutation of $X$ as

$$\mathcal{S}_X = \{f : X \to X \mid f \text{ is bijective}\}.$$

With the composition operation form the *symmetric group* or the *permutation group* of $X$. Particular attention is devoted to the case $X = [n]$, we write $S_n = \mathcal{S}_X$ and it called *symmetric group* or the *permutation group* of $n$ elements.

- Let $\mathbb{Z}_n$ be the group of integers modulo $n$ with the addition operation, they are called *finite cyclic groups* of order $n$.
- Given two groups $G$ and $H$, the direct product $G \times H$ of them is still a group. The set of the elements is the Cartesian product of $G$ and $H$ while the sum is defined as

$$(g_1, h_1) \cdot_{G\times H} (g_2, h_2) = (g_1 \cdot_G g_2, h_1 \cdot_H h_2).$$

Now, we introduce the notion of group homomorphism, a transformation between groups which preserves the operation.

**Definition 2.1.3.** A *group homomorphism* is a map

$$\phi : G \to H$$

between $G$ and $H$ groups such that, for each $g, h \in G$

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

A bijective group homomorphism is called *isomorphism*.

**Example 2.1.4.** Here we present some fundamental examples of group homomorphisms.

- The map $\phi : \mathbb{Q} \to \mathbb{Q}$ defined by $\phi(x) = 2x$ for each $x \in \mathbb{Q}$ is an isomorphism of additive groups. Note that the restriction $\phi_{|\,\mathbb{Z}}$ is a homomorphism on $\mathbb{Z}$ but it is not an isomorphism since it is not surjective.
- The map $\exp : \mathbb{R} \to \mathbb{R}_{>0}$ is an isomorphism between the additive group $(\mathbb{R}, +)$ and the multiplicative group $(\mathbb{R}_{>0}, \cdot)$.

**Definition 2.1.5** (Cosets)**.** Let $G$ be a group and $H$ be a subgroup of $G$. The *set of left cosets* of $G$ by $H$ is the set $G/H = \{gH \mid g \in G\}$, where $gH = \{gh \mid h \in H\}$ are the *left cosets* of $H$. Similarly, we define the set of *right cosets* as $H\backslash G = \{Hg \mid g \in G\}$. Let $K$ be a second subgroup of $G$, we define the *double coset* of $H$ and $K$ with respect to an element $g \in G$ as the set $HgK = \{hgk \mid h \in H, k \in K\}$. The set of double cosets is denoted as $H\backslash G/K$.

**Example 2.1.6.** Relevant examples of left cosets include the following:

1. Consider $G = \mathbb{Z}$ and the subgroup $H = n\,\mathbb{Z}$ of integers multiples of $n$. The quotient $G/H$ is a group and is isomorphic to the cyclic group of $n$ elements, $\mathbb{Z}_n$.

2. Consider $G = S_3$, the symmetric group on three elements, and the subgroup $H = \{(1), (12)\}$. The quotient $G/H$ is a group and is isomorphic to $S_2$, symmetric group on two elements.

3. Consider $G = S_n$, the symmetric group on $n$ elements, and the subgroup $H = A_n$, the alternating group on $n$ elements. The quotient $G/H$ is a group and is isomorphic to $\mathbb{Z}_2$.

A result that will play a central role in Chapter 3 is the characterization of closed multiplicative subgroups of $\mathbb{R}^*$, which we state and prove here.

**Proposition 2.1.7.** *The closed multiplicative subgroups of $\mathbb{R}^*$ are as follows:*

- *The trivial subgroup $\langle 1\rangle$,*
- *The order-two subgroup $\langle \pm 1\rangle$,*
- *Discrete positive subgroup $\langle b\rangle = \{b^n \mid n \in \mathbb{Z}\}$ for some $b > 1$,*
- *Discrete subgroups $\langle \pm b\rangle = \{\pm b^n \mid n \in \mathbb{Z}\}$ for some $b > 1$,*
- *The subgroup of all positive reals $\mathbb{R}_{>0}$,*
- *The group $\mathbb{R}^*$ itself.*

*Proof.* Multiplicative subgroups of $\mathbb{R}_{>0}$ and additive subgroups of $\mathbb{R}$ are linked by the exponential map which is an isomorphism. Additive subgroups of $\mathbb{R}$ are divided into three different type: finite ($\langle 0\rangle$), unbounded and discrete ($\alpha\,\mathbb{Z}$ for each $\alpha \in \mathbb{R}_{>0}$) and dense. They map through into $\mathbb{R}_{>0}$ as finite ($\langle 1\rangle$), unbounded and discrete ($\{b^n = e^{n\alpha}\}_{n\in\mathbb{Z}} \cong \mathbb{Z}$) and dense. We can get the multiplicative subgroups of $\mathbb{R}^*$ noticing that $\mathbb{R}^* \cong \mathbb{Z}_2 \times \mathbb{R}_{>0}$. □

## 2.2 Group Actions and Equivariant Maps

Let $G$ be a group and $X$ be a set. An *action* of the group $G$ on the set $X$ is a function

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

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

- For the identity element $e$ in $G$, the *identity condition* $\phi_e = id_X$ holds.
- For all $g, h \in G$, the *compatibility condition* $\phi_g \circ \phi_h = \phi_{gh}$ holds.

In this context, we often write $g \cdot x$ or simply $gx$ instead of $\phi_g(x)$. A *$G$-set* is a set $X$ equipped with a group action of $G$. This means that there is a well-defined action $\cdot : G \times X \to X$ satisfying the properties of a group action as described above.

**Example 2.2.1.** We present some examples of group actions.

(i) Any group $(G, \cdot)$ acts on itself by left multiplication. That is, for each $g \in G$, define $\phi_g : G \to G$ by $\phi_g(x) = g \cdot x$ for each $x \in G$.

(ii) Let $\mathbb{K}$ be one of $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, or $\mathbb{H}$. The additive group $(\mathbb{K}, +)$ acts on $\mathbb{K}$ by translations; explicitly, for each $g \in \mathbb{K}$ define $\phi_g : \mathbb{K} \to \mathbb{K}$ by $\phi_g(x) = x + g$ for all $x \in \mathbb{K}$. This defines a left action of $\mathbb{K}$ on itself.

(iii) Let $\mathbb{K}$ be one of $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, or $\mathbb{H}$. The multiplicative group $\mathbb{K}^*$ acts on $\mathbb{K}$ by scaling; explicitly, for each $g \in \mathbb{K}^*$ define $\phi_g : \mathbb{K} \to \mathbb{K}$ by $\phi_g(x) = gx$ for all $x \in \mathbb{K}$. This defines a left action of $\mathbb{K}^*$ on $\mathbb{K}$.

(iv) The symmetric group $S_n$ acts on $[n]$ by $\sigma \cdot i \coloneqq \sigma(i)$.

(v) The group $\mathrm{GL}_n(\mathbb{K})$ acts on $\mathbb{K}^n$ by $A \cdot x \coloneqq Ax$.

(vi) If $G$ acts on $X$, then $G$ acts on the space $\mathbb{R}^X$ of real-valued functions on $X$ by

$$(g \cdot f)(x) \coloneqq f(g^{-1}x).$$

Throughout the following sections, it will often be convenient to decompose $G$-sets into a disjoint union of subsets, each minimal (in a sense specified in Definition 2.2.2) and equipped with a compatible $G$-action.

**Definition 2.2.2.** Let $G$ be a group acting on a set $X$. An *orbit* in $X$ is a subset $Y \subseteq X$ such that for each $x \in Y$, we have $Y = \{g \cdot x \mid g \in G\}$. The set $X$ can be decomposed into a disjoint union of orbits under the action of $G$. This is called the *orbit decomposition* of $X$, and if $X$ is finite, the decomposition can be written as

$$X = X_1 \sqcup \cdots \sqcup X_n,$$

where $X_1, \ldots, X_n$ are the distinct orbits of $X$. If an action presents only a single orbit, we say that the action is *transitive*.

**Example 2.2.3.** Let $\mathbb{R}$ act on $\mathbb{R}$ by addition. This action is transitive. On the other hand, if $\mathbb{R}^*$ acts on $\mathbb{R}$ by multiplication, then the orbit decomposition is

$$\mathbb{R} = \{0\} \sqcup \mathbb{R}^*.$$

Another fundamental concept for our treatment is that of a function between $G$-sets that preserves actions, which is more formally specified in Definition 2.2.4.

**Definition 2.2.4.** Let $X$ and $Y$ be two $G$-sets, a map $f : X \to Y$ is *$G$-equivariant* if

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

for each $x$ in $X$.

For each $T < \mathbb{R}^*$ in Proposition 2.1.7, acting on $\mathbb{R}$ by multiplication, we characterize the set of $T$-equivariant continuous functions, namely

$$\mathcal{F}_T := \{f \in \mathcal{C}(\mathbb{R}, \mathbb{R}) \mid f(gx) = gf(x) \text{ for all } g \in T,\ x \in \mathbb{R}\}\,.$$

To exhibit a wide variety of functions of this kind, we now state and prove Proposition 2.2.5, which gives a complete characterization of $T$-equivariant continuous functions, including nonlinear examples.

**Proposition 2.2.5.** *Let $T \leq \mathbb{R}^*$ be a closed multiplicative subgroup as in Proposition 2.1.7. For each such $T$, the space $\mathcal{F}_T$ of $T$-equivariant continuous functions is characterized as follows.*

- *If $T = \langle 1 \rangle$, then $\mathcal{F}_T = \mathcal{C}(\mathbb{R}, \mathbb{R})$.*
- *If $T = \langle \pm 1 \rangle$, then $\mathcal{F}_T = \mathcal{O}(\mathbb{R})$, the space of odd continuous functions.*
- *If $T = \langle b \rangle = \{b^n \mid n \in \mathbb{Z}\}$ for some $b > 1$, then $\mathcal{F}_T$ consists of all continuous functions of the form*

$$f(x) = \begin{cases} f_{\eta_+}(x) & x > 0 \\ 0 & x = 0 \\ f_{\eta_-}(-x) & x < 0 \end{cases}. \tag{2.1}$$

  *Here $\eta_\pm \in \mathcal{C}([1, b], \mathbb{R})$ satisfy $\eta_\pm(b) = b\,\eta_\pm(1)$, and $f_{\eta_\pm} : \mathbb{R}_{>0} \to \mathbb{R}$ is defined by*

$$f_{\eta_\pm}(x) := b^n\, \eta_\pm\Big(\frac{x}{b^n}\Big)\,,$$

  *where $n$ is the unique integer such that $\frac{x}{b^n} \in [1, b)$, see Lemma 2.2.6 for the well-definedness of this construction.*

- *If $T = \langle -b \rangle = \{\pm b^n \mid n \in \mathbb{Z}\}$ for some $b > 1$, then $\mathcal{F}_T$ consists of all functions as in the previous case, with the additional constraint $\eta_+ = \eta_-$.*
- *If $T = \mathbb{R}_{>0}$, then $\mathcal{F}_T$ consists of all two-sided linear functions, i.e., functions of the form $f(x) = a_+ x$ for $x \geq 0$ and $f(x) = a_- x$ for $x \leq 0$.*
- *More generally, if $T = \mathbb{R}^*$, then $\mathcal{F}_T$ consists of all linear functions on $\mathbb{R}$, i.e., $f(x) = ax$.*

The following Lemma 2.2.6 implies that for each $x \in \mathbb{R}_{>0}$ there exist a unique $n \in \mathbb{Z}$ such that $\frac{x}{b^n} \in [1, b)$. This observation ensures that Equation 2.1 is well-defined.

**Lemma 2.2.6.** *Let $b$ a real number, $b > 1$, and $T = \langle b^n \rangle_{n \in \mathbb{Z}}$. For each positive real number $x$ there exist a unique decomposition $x = b^n y$ such that $y \in [1, b)$ and $n \in \mathbb{Z}$.*

*Proof.* Choose $n$ such that $x \in [b^n, b^{n+1})$, we have that $x = b^n y$ where we define $y = \frac{x}{b^n} \in [1, b)$. Sets $\{[b^n, b^{n+1})\}_{n\in\mathbb{Z}}$ are a collection of disjoint intervals covering $\mathbb{R}$. Hence $x$ belongs only to one of those, say $x \in [b^n, b^{n+1})$, we have the uniqueness of the decomposition $x = b^n y$ such that $y \in [1, b)$. □

The following technical lemmas will be necessary to prove Proposition 2.2.5.

**Lemma 2.2.7.** *Let $b > 1$ be a real number, and let $T = \langle b^n \rangle_{n\in\mathbb{Z}}$. A continuous function $f$ is $T$-equivariant if and only if there exist two continuous functions $\eta_\pm : [1, b] \to \mathbb{R}$ such that $\eta_\pm(b) = b\eta_\pm(1)$ and*

$$f(x) = \begin{cases} f_{\eta_+}(x) & x > 0 \\ 0 & x = 0 \\ f_{\eta_-}(-x) & x < 0 \end{cases}.$$

*Proof.* ($\Rightarrow$) Define $\eta_\pm = f_{|[\pm 1, \pm b]}$. Multiplicativity and continuity of $f$ implies all the required properties of $\eta_\pm$.
($\Leftarrow$) We only need to prove that $f$ constructed in this way is continuous on $\mathbb{R}$. We will prove the continuity of $f$ on $\mathbb{R}_{>0}$ and $\mathbb{R}_{<0}$, and in 0. The two former cases boil down to proving continuity of $f_{\eta_\pm}$ respectively. As those proofs are analogous, we only present the one for $f_{\eta_+}$. Note that for each $x \in \mathbb{R}_{>0} \setminus T$, there exists an interval $x \in I \subseteq \mathbb{R}_{>0} \setminus T$ and an integer $n$ such that $f_{\eta_+}(x) = b^n \eta(\frac{x}{b^n})$ for each $x \in I$. Hence $f_{\eta_+}$ is continuous on $\mathbb{R}_{>0} \setminus T$. Now see that

$$\lim_{x\to b^{n+}} f_{\eta_+}(x) = \lim_{x\to 1^+} b^n \eta(x) = \lim_{x\to b^-} b^{n-1}\eta(x) = \lim_{x\to b^{n-}} f_{\eta_+}(x)$$

because $\eta_+(b) = b\eta_+(1)$. It remains to prove continuity of $f$ in 0, i.e.,

$$\lim_{x\to 0^-} f_{\eta_+}(x) = \lim_{x\to 0^+} f_{\eta_-}(x).$$

Note that $\eta_+$ is continuous and limited as it is defined on a compact subset of $\mathbb{R}$, hence we can define $m = \min_{x\in[1,b]} \eta_+(x)$ and $M = \max_{x\in[1,b]} \eta_+(x)$, note that $f_{m\cdot 1_{[1,b]}} \leq f_{\eta_+} \leq f_{M\cdot 1_{[1,b]}}$ where $1_{[1,b]}$ is the constant function with value 1 defined on $[1, b]$. It is easy to see that $\lim_{x\to 0^+} f_{m\cdot 1_{[1,b]}} = \lim_{x\to 0^+} f_{M\cdot 1_{[1,b]}} = 0$. □

We now state Lemma 2.2.8 whose proof will essentially contain the proof of Proposition 2.2.5.

**Lemma 2.2.8.** *Each $T$-equivariant continuous function is linear if and only if $T$ is dense in $\mathbb{R}$.*

*Proof.* Note that if $f$ is $T$-equivariant, by definition, $f(t) = tf(1)$ for each $t \in T$. As $T$ is dense in $\mathbb{R}$, $f$ is linear on its entire domain $\mathbb{R}$ by continuous extension. If $T$ is dense in $\mathbb{R}_{>0}$, $f(t) = tf(1)$ and $f(-t) = tf(-1)$ for each $t > 0$, hence $f$ is semilinear. On the other hand, if $T$ is not dense, suppose $T$ is $\langle 1\rangle$ or $\langle \pm 1\rangle$, then $\mathcal{F}_T$ are $\mathcal{C}(\mathbb{R})$ and odd functions respectively, which contain non-linear functions.

If $T = \langle b^n \rangle_{n\in\mathbb{Z}}$. Let us now proceed with the construction of a function that is continuous, $T$-equivariant, and non-linear. Let $\eta_\pm : [1, b] \to \mathbb{R}$ be two continuous bump function and let $f$ be a continuous function as defined in Equation 2.1, this function is multiplicative and non-linear. This concludes the proof as we have listed all the possible cases of $T$ presented in Proposition 2.1.7. □

*Proof of Proposition 2.2.5.* Thanks to Propositions 2.1.7 and 2.2.7, the presented proof of Lemma 2.2.8 gives a complete characterization of $T$-equivariant functions with respect to multiplicative subgroups $T$ of $\mathbb{R}$. □

## 2.3 Representation Theory

**Definition 2.3.1.** A *representation* of a group $G$ on a vector space $V$ is a group homomorphism

$$\rho : G \to \mathrm{GL}(V).$$

Equivalently, $\rho$ is a left action of $G$ on $V$ such that $\rho(g)$ is linear for each $g \in G$.

**Example 2.3.2.** Consider $G = S_n$.

- $\dim V = 1$ and $\rho(g) = \mathrm{id}$ for each $g \in S_n$. This is called the *trivial* representation of $S_n$.
- $\dim V = 1$ and $\rho(g) = \mathrm{sgn}(g) \cdot \mathrm{id}$. This is called the *sign* representation of $S_n$.

**Definition 2.3.3.** Let $T \leq \mathbb{R}^*$ be a subgroup. We say that a representation $\rho : G \to \mathrm{GL}_n(\mathbb{R})$ is

- *Nonnegative* if all entries of $\rho(g)$ are nonnegative for every $g \in G$.
- *$T$-monomial* if, for every $g \in G$, each row and each column of $\rho(g)$ has exactly one nonzero entry, and that entry belongs to $T$. Such matrices are called *$T$-monomial matrices*, and we denote by $\mathcal{M}(T)$ the set of $T$-monomial matrices.
- A *permutation representation* if it is $\{1\}$-monomial, equivalently, if each $\rho(g)$ is a permutation matrix. We discuss this class of representations in more detail in Section 4.1.
- A *signed permutation representation* if it is $\{\pm 1\}$-monomial, equivalently, if each $\rho(g)$ is a signed permutation matrix.

**Example 2.3.4.** The following matrices $P, S$ and $M$ are respectively examples of permutation matrices, signed permutation matrices and $\langle 2\rangle$-monomial matrices:

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad S = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 2 \\ 2 & 0 & 0 \end{bmatrix}.$$

In particular, we will be interested in the case of continuous representations.

**Definition 2.3.5** (Representations of Topological Groups)**.** Let $G$ be a topological group and $V$ be real vector space. A *continuous representation* of $G$ on $V$ is a continuous group homomorphism $\rho : G \to \mathrm{GL}(V)$.

**Example 2.3.6.** Consider the general linear group $\mathrm{GL}(V)$ of invertible $n \times n$ matrices, with the topology induced by the Euclidean norm on matrices. Let $V = \mathbb{R}^n$ be the standard Euclidean space. The map $\rho : G \to \mathrm{GL}(V)$ defined by $\rho(A)(v) = Av$ is a continuous representation. Similarly, restricting this representation to $\mathrm{SO}(n)$, we get a representation of $\mathrm{SO}(n)$.

The concept of equivariance can be adapted to the case of linear representations as follows.

**Definition 2.3.7.** A linear map $\Phi : V \to W$ is *$G$-equivariant* with respect to the representations $\rho_1 : G \to \mathrm{GL}(V)$ and $\rho_2 : G \to \mathrm{GL}(W)$ if

$$\rho_2 \circ \Phi = \Phi \circ \rho_1.$$

We will denote the space of all $G$-equivariant maps between $\rho_1$ and $\rho_2$ by $\mathrm{Hom}_G(\rho_1, \rho_2)$ or $\mathrm{Hom}_G(V, W)$ when $\rho_1$ and $\rho_2$ will be clear from the context.

**Definition 2.3.8.** A representation $\rho : G \to \mathrm{GL}(V)$ is *irreducible* if there exists no non-trivial subspace $W$ of $V$ such that $\rho(g)(W) \subseteq W$ for each $g \in G$.

The following fundamental result in representation theory of finite groups states that there is always a decomposition into irreducible representations.

**Theorem 2.3.9.** *Let $G$ be a finite group and $V$ a finite-dimensional representation of $G$. Then, there exists a decomposition*

$$V = V_1 \oplus \cdots \oplus V_m$$

*where each $V_i$ is irreducible for $\rho$. This decomposition is unique up to isomorphism and permutation of the factors.*

We will also often employ the following fundamental result.

**Lemma 2.3.10** (Schur's Lemma)**.** *Let $V$ and $W$ be non-isomorphic irreducible representations, there is only one $G$-equivariant linear map between them and it is the trivial one.*

**Definition 2.3.11.** The fixed set for a representation $\rho : G \to \mathrm{GL}(V)$ is defined as

$$V^G := \{v : gv = v\}.$$

Note that $V^G$ is a representation for $G$ and the action is trivial.

**Theorem 2.3.12.** *The following properties are true for representations of a finite group $G$.*

- $\dim V^G$ *is the multiplicity of the trivial representation in* $V$*,*
- $\dim \mathrm{Hom}_G(V, W) = \dim(V \otimes W)^G$.

*Remark* 2.3.13. Let $V \otimes W$ be a finite-dimensional $G \times H$-representations and $V_i$'s a complete list of irreducible $G$-representations and $W_j$'s a complete list of irreducible $H$-representations, then $V_i \otimes W_j$ 's is a complete list of irreducible $(G \times H)$-representations (See [101] for proofs in case $G$ and $H$ are finite). If $m_i$ is the multiplicity of $V_i$ in $V$ and $n_j$ is the multiplicity of $W_j$ in $W$ then the multiplicity of $V_i \otimes W_j$ in $V \otimes W$ is $m_i n_j$. This can be easily seen by writing the irreducible decompositions of $V$ and $W$ and use the distributive property of tensor products and direct sums. Note that the same is true if $G$ and $H$ are compact groups and the representations are continuous. If $S$ is an $G$-isotypic component of $V \times W$ of type $\rho$ then it is $H$-invariant and each $h \in H$ acts as $G$-equivariant endomorphism of $S$. By Lemma 2.3.10, $S = \rho \otimes \sigma$, which is $\bigoplus_i \rho \otimes \sigma_i$, where $\sigma_i$ are irreducible $H$-representations. Iterating the decomposition if necessary and by the finite dimension of $V$ and $W$, we conclude.

## 2.4 Affine Equivariant Maps

We will work with neural networks composed of affine layers, and we provide a brief introduction to this setting here.

**Definition 2.4.1.** Let $V$ and $W$ be two $\mathbb{K}$-vector spaces and define the translation of a vector $w$ in $W$ as a non-linear bijective map

$$\tau_w : v \mapsto v + w.$$

Define the space of *affine maps* from $V$ to $W$ as

$$\mathrm{Aff}(V, W) = \{\tau_w \circ f | w \in W \text{ and } f \in \mathrm{Hom}(V, W)\}.$$

Note that is a more general definition with respect to the standard one, where $f$ is an isomorphism of a vector space $V$.

**Theorem 2.4.2.** *The decomposition of an affine map* $\phi \in \mathrm{Aff}(V, W)$ *in translational part* $\tau_w$ *and* $f$ *linear part is unique.*

*Proof.*

$$\tau_{w_1} \circ f_1 = \phi = \tau_{w_2} \circ f_2,$$

evaluating in 0 leads to

$$w_1 = \phi(0) = w_2.$$

Write $w = w_1 = w_2$, and note that

$$\tau_w \circ f_1 = \tau_w \circ f_2,$$

by the bijectivity of translations,

$$f_1 = f_2.$$

□

Let $V$ and $W$ be $G$-representation. An affine map $\phi$ in $\mathrm{Aff}(V,W)$ is $G$-equivariant if $\phi \circ g = g \circ \phi$ for each $g \in G$, write the set of $G$-equivariant affine maps from $V$ to $W$ as $\mathrm{Aff}_G(V,W)$.

**Theorem 2.4.3.** *A map $\phi = \tau_w \circ f \in \mathrm{Aff}_G(V,W)$ if and if $f \in \mathrm{Hom}_G(V,W)$ and $v$ is invariant.*

*Proof.* Note that for each $g \in G$,

$$g \circ \tau_w \circ f = \tau_{gw} \circ (g \circ f).$$

Observe that

$$\phi \circ g = g \circ \phi,$$

if and only if

$$\tau_w \circ f \circ g = g \circ \tau_w \circ f = \tau_{gw} \circ (g \circ f)$$

if and only if, by the previous proposition,

$$w = gw$$

and

$$f \circ g = g \circ f$$

for each $g \in G$. □

We can define restrictions of maps $\theta_G$, $\lambda_G$, and $\tau_G$ as follows

$$\theta_G : \begin{array}{c} \mathrm{Hom}_G(V,W) \oplus W^G \to \mathrm{Aff}_G(V,W) \\ (\phi, v) \mapsto \tau_v \phi \end{array}$$

$$\lambda_G : \begin{array}{c} \mathrm{Aff}_G(V,W) \to \mathrm{Hom}_G(V,W) \\ f \mapsto f - f(0) \end{array} \quad \text{and} \quad \tau_G : \begin{array}{c} \mathrm{Aff}_G(V,W) \to W^G \\ f \mapsto f(0). \end{array}$$

Theorem 2.4.3 implies that $\theta_G$ is an isomorphism and a similar proof to the one of Theorem 2.4.2 shows that both $\lambda_G$ and $\tau_G$ are linear, equivariant, and surjective. When it will be clear we are working in the equivariant setting, we will drop the subscript and just write $\theta$, $\lambda$, and $\tau$.

In the main text we will often use the following results.

**Proposition 2.4.4.** *Let $V_1$, $V_2$, and $V$ be $G$-representations. Then,*

$$\mathrm{Hom}_G(V_1 \oplus V_2, V) = \mathrm{Hom}_G(V_1, V) \oplus \mathrm{Hom}_G(V_2, V)$$

*and*

$$\mathrm{Hom}_G(V, V_1 \oplus V_2) = \mathrm{Hom}_G(V, V_1) \oplus \mathrm{Hom}_G(V, V_2).$$

**Proposition 2.4.5.** *Let $V$ and $W$ be $G$-representations, and let $G$ act trivially on $\mathbb{R}^n$ and $\mathbb{R}^m$. Then,*

$$\mathrm{Hom}_G(V \otimes \mathbb{R}^n, W \otimes \mathbb{R}^m) \cong \mathrm{Hom}_G(V,W) \otimes \mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^m),$$

*and*

$$(V \otimes \mathbb{R}^n)^G \cong V^G \otimes \mathbb{R}^n \,.$$

*In other words, recalling that* $\mathrm{Aff}_G(V,W) \cong \mathrm{Hom}_G(V,W) \oplus W^G$ *and since* $\mathrm{Hom}(\mathbb{R}^n, \mathbb{R}^m) \cong \mathbb{R}^{n\times m}$ *is the set of* $n \times m$ *matrices over* $\mathbb{R}$*, understanding the structure of* $\mathrm{Aff}_G(V \otimes \mathbb{R}^n, W \otimes \mathbb{R}^m)$ *reduces to understanding the structure of* $\mathrm{Aff}_G(V,W)$*.*

## 2.5 Equivariant Neural Networks

We now define equivariant neural networks, which were first introduced by [121] as *Group Representation Networks* and by [27] as *G-Steerable Convolutional Networks*. The same model later appears in other papers such as [8] under the name of *equivariant neural networks*, which is the notation we adopt in this paper.

Given a group $G$ and arbitrary $G$-representations $V_i$ for $1 \leq i \leq m$, a $G$-equivariant neural network is a composition

$$\Phi = \phi_m \circ \tilde{f}_{m-1} \circ \phi_{m-1} \circ \cdots \circ \tilde{f}_1 \circ \phi_0, \tag{2.2}$$

where each *activation* $\tilde{f}_i : V_i \to V_i$ is a $G$-equivariant function, and $\phi_i \in \mathrm{Aff}_G(V_i, V_{i+1})$ is an affine $G$-equivariant map.

An activation $\tilde{f}_i : V_i \to V_i$ is *point-wise* if there exist a basis $\mathcal{B}_i = \{v_1, \ldots, v_m\}$ of $V_i$ and a real scalar function $f$ such that

$$\tilde{f}_i(a_1 v_1 + \cdots + a_m v_m) = f_i(a_1)v_1 + \cdots + f_i(a_m)v_m \;\; \forall\, a_1, \ldots, a_m \in \mathbb{R}\,.$$

In this case, we say that $f_i$ induces $\tilde{f}_i$ on $\mathcal{B}$. Note that we do not constrain activations to be non-linear or non-affine but, from now on, we only consider continuous functions $\mathcal{C}(\mathbb{R})$. This condition is not particularly restrictive as continuous functions constitute a wide class of function which includes all commonly employed point-wise activations (such as ReLU, tanh,...), they are compatible with backpropagation, and strictly encompasses activations dealt by [121].

# Chapter 3

# Equivariant Activations

In this chapter, using tools from representation theory [39] and matrix group theory [36], we present two general characterizations of pairs of representations and real-valued functions that induce equivariant activations. In more details, we consider classes $\mathcal{F}$ of activation functions and groups $\mathcal{M}$ of representation matrices, and we investigate when they can be combined to obtain an equivariant layer. We start by defining operations that map any $\mathcal{F}$ to a corresponding maximal admissible group of matrices $\mathcal{M}(\mathcal{F})$, and vice-versa map $\mathcal{M}$ to a maximal admissible family of activations $\mathcal{F}(\mathcal{M})$. We highlight the dual nature of those operations and how their composition stabilizes, leading to a finite family of maximal classes. Finally, for these few maximal classes, we exhibit the dual pairings $(\mathcal{F}, \mathcal{M})$, thereby obtaining a complete classification of admissible activation–representation pairs, which includes the classification of [121] as a special case. We also characterize the pairs $(\mathcal{F}, \mathcal{M})$ up to isomorphism of the representations induced by $\mathcal{M}$.

We then explore relevant implications these theorems. First, we consider disentangled equivariant networks [27], and we show that they admit point-wise activations if and only if the linear representations underlying their feature spaces are trivial (Section 3.3). Second, we specialize our results to particular cases of equivariant networks of practical relevance (Section 3.4). For rotation-equivariant networks [119], we show that all admissible networks are invariant, and hence unable to learn many equivariant tasks such as segmentation or detection. This proves a barrier for the use of equivariant networks with point-wise activations in this kind of tasks. Instead, for Invariant Graph Networks (IGNs) [78] and geometric IGNs [28, 34, 54], including the $k$-order invariant and equivariant networks presented by [78], we show that the choice of admissible representation layers compatible with point-wise activations is not only limited to those described by [78], and in fact we highlight and fully characterize a wider class of permutation equivariant networks in terms of subgroups of the symmetric group.

In brief, our contributions are summarized as follows:

- We provide a characterization theorem for equivariant neural networks with continuous point-wise activations, by showing the existence of a finite number of maximal sets of equivariant classes, enumerating them, and providing an explicit dual pairing between activations and representations,
- We use this result to give an exhaustive description of networks that are equivariant with respect to finite groups, and to show a barrier in the use of disentangled equivariant networks,
- We discuss implications of this theorem in practical and relevant scenarios, namely highlighting a severe limitation of rotation-equivariant networks.

## 3.1 Characterization Theorems

For our construction we will need also the following relations between activations and representation matrices. Given a group $G$ and a representation

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

of $G$, we consider a point-wise activation

$$\tilde{f} : V \to V \quad \text{induced by} \quad f : \mathbb{R} \to \mathbb{R}$$

on a basis $\mathcal{B}$ of $V$. The image of $\rho$ in $\mathrm{GL}(V)$ with respect to the basis $\mathcal{B}$ forms a group of matrices which we denote $\mathcal{M}$, and note that $\tilde{f}$ is $G$-equivariant if and only if it commutes with respect to the matrices in $\mathcal{M}$, i.e., $\tilde{f}(Mv) = M\tilde{f}(v)$ for each $M \in \mathcal{M}$ and $v \in V$, where both $v$ and $M$ are written on the basis $\mathcal{B}$.

This paper proposes a simple procedure to recover the maximal group of representation matrices $\mathcal{M}$ compatible with a given class $\mathcal{F}$ of activation functions, and vice-versa the widest class of functions $\mathcal{F}$ given a group of representation matrices $\mathcal{M}$. The precise definition is as follows.

**Definition 3.1.1.** Given a family of activations $\mathcal{F} \subseteq \mathcal{C}(\mathbb{R})$, we define *the maximal group* $\mathcal{M}(\mathcal{F})$ *admitted by* $\mathcal{F}$ as the set of all matrices in $\mathrm{GL}_n(\mathbb{R})$ which commutes with each $\tilde{f} : V \to V$ induced by $f \in \mathcal{F}$. Conversely, given $\mathcal{M} \subseteq \mathrm{GL}_n(\mathbb{R})$, *the maximal set* $\mathcal{F}(\mathcal{M})$ *admitted by* $\mathcal{M}$ is the set of all functions $f \in \mathcal{C}(\mathbb{R})$ such that $\tilde{f}$ induced by $f$ commutes with each $M \in \mathcal{M}$.

Observe that $\mathcal{M}(\mathcal{F})$ has trivially a group structure with respect to matrix product, since if $M_1, M_2 \in \mathcal{M}(\mathcal{F})$ commute with $\tilde{f}$, then also their product $M_1 M_2$ does. Note that some groups of matrices or families of functions are contained in each other. For example, permutation matrices are contained signed permutation matrices, while $\langle b^2 \rangle$-monomial matrices are contained into

$\langle b \rangle$-monomial matrices, but they are not maximal following Definition 3.1.1. Moreover, the following stabilization lemma (Lemma 3.1.2) proves that the maps to the maximal sets stabilize after two iterations for any initial choice of $\mathcal{M}$ or $\mathcal{F}$, proving the duality of the operators $\mathcal{M}(\cdot)$ and $\mathcal{F}(\cdot)$ on maximal elements. See Section D.1.2 for the proof.

**Lemma 3.1.2.** *The group of matrices $\mathcal{M}' = \mathcal{M}(\mathcal{F}(\mathcal{M}))$ is the largest group in $\mathrm{GL}_n(\mathbb{R})$ for which $\mathcal{F}(\mathcal{M}') = \mathcal{F}(\mathcal{M})$, and $\mathcal{F}' = \mathcal{F}(\mathcal{M}(\mathcal{F}))$ is the largest family of functions in $\mathcal{C}(\mathbb{R})$ for which $\mathcal{M}(\mathcal{F}') = \mathcal{M}(\mathcal{F})$.*

Thanks to Lemma 3.1.2, it is sufficient to provide explicit pairings between $\mathcal{M}$ and $\mathcal{F}(\mathcal{M})$ (or $\mathcal{F}$ and $\mathcal{M}(\mathcal{F})$) just for maximal admissible groups $\mathcal{M}$, and similarly for maximal admissible families of functions $\mathcal{F}$. Indeed, given any other $\overline{\mathcal{M}}$ (or $\overline{\mathcal{F}}$) which is not maximal, it is sufficient to find a superset $\mathcal{M}$ of $\overline{\mathcal{M}}$ which is a maximal group in the sense of Definition 3.1.1, and this will give a set of admissible functions $\mathcal{F}$ which are equivariant also for $\overline{\mathcal{M}}$.

We can now state our main result, which shows that, under mild assumptions, these maximal groups (or, equivalently, maximal function classes) are only a very limited number. Moreover, for each of these we provide the exact correspondence between $\mathcal{M}$ and $\mathcal{F}$, thus providing an exhaustive classification of all possible admissible pairs of $\mathcal{M}$ and $\mathcal{F}$. Since the set $\mathcal{F}$ represents the activation functions, we make the natural assumptions that it does not contain only affine functions.

**Theorem 3.1.3.** *Let $\mathcal{F}$ be a family of non-affine real functions. Maximal admissible pairs $(\mathcal{M}, \mathcal{F})$ are of the type*

$$(\mathcal{M}(T), \mathcal{F}_T)$$

*for some closed multiplicative group $T < \mathbb{R}^*$, $\mathcal{M}(T)$ are $T$-monomial matrices, and $\mathcal{F}_T$ are $T$-equivariant functions. Recall Definition 2.3.3 and (2.2).*

We classified groups of matrices and families of activation functions commuting with each others, and this classification is exhaustive. However, for compact groups $G$ we show that there is another tool to enlarge the set of possible activation functions, namely moving to isomorphic representation which admit a wider family of activation functions. Indeed, note that the representation $\rho : \mathbb{Z}_2 \to \mathrm{GL}_2(\mathbb{R})$ given by

$$\rho(0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \rho(1) = \begin{bmatrix} 0 & 2 \\ \frac{1}{2} & 0 \end{bmatrix}$$

We have shown in Theorem 3.1.3 that in fact the broader set of $\langle 2 \rangle$-equivariant functions is admissible. Despite $\langle 2 \rangle$-equivariant functions being a large set, they still do not include common activations such as tanh, sigmoid, or softplus. However, by a basis change through the matrix $B = \mathrm{diag}(1/\sqrt{2}, \sqrt{2})$, we obtain the isomorphic representation

$$B\rho(0)B^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B\rho(1)B^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which is the standard permutation representation of $\mathbb{Z}_2$, and thus commutes with all continuous functions (see Theorem 3.1.3). This approach is generalized to the case of arbitrary compact groups by the following theorem, showing that in this case representations are always isomorphic to (signed) permutation representations, which have the widest admissible family of activations. We refer to Section D.1.2 for its proof.

**Theorem 3.1.4.** *Let $G$ be a compact group and assume that $\mathcal{F}$ is not a set of only affine functions. Then any representation of $G$ admitted by $\mathcal{F}$ is isomorphic to a subgroup of the (signed) permutation matrices. Thus, $\mathcal{F}$ can always be chosen as the entire the set of (odd) continuous functions.*

Next, we will delve into the practical implications of Theorems 3.1.3 and 3.1.4, focusing on non-odd non-affine activations, the most used activations in practice. We further specialize these results to the case of finite groups, where the only non-trivial admissible representations is induced by permutation representations, which can be described with more precision. We explore the implications of these results on neural networks designed for processing first-order relational structures like sets and point clouds [95, 124]. These networks have proven highly effective in practice, showcasing efficiency and accuracy when dealing with such unordered data. Finally, in this setting, we show that admissible disentangled representations coupled with point-wise activations are trivial.

## 3.2 Finite Groups

As the majority of activations used in practice are induced by non-odd non-linear functions, we present a corollary of Theorem 3.1.4 that completely describes representations for finite groups that can be used in these cases. We show that admissible representations are only permutations representations up to positive-scaling of the basis.

**Corollary 3.2.1.** *Let $G$ be a finite group, and let $f : V \to V$ be a non-odd non-affine equivariant activation function on a $G$-representation $V$ defined on the basis $\mathcal{B}$. Then there exists a positive scaling of $\mathcal{B}$ and a collection of subgroups $H_i < G$ such that $V = \mathbb{R}^{G/H_1} \times \cdots \times \mathbb{R}^{G/H_m}$.*

*Proof.* We can consider $V$ to be a finite permutation representation thanks to Theorem 3.1.4, as the only admissible bounded matrix groups are permutation matrices, while signed permutation matrices are compatible only with odd functions. A permutation representation has an underlying permutation set that can be decomposed in the disjoint union of orbits. Recall that, for any $X$ on which $G$ acts transitively, there exists a set bijection $X \cong G/H$ for some subgroup $H$ of $G$. Hence, a permutation representation $V$ can be decomposed into the direct sum $V = \mathbb{R}^{\coprod G/H_i}$ for a collection of finite index subgroups $H_i < G$ such that there is a bijection between the orbits $X_i$ of $G$ in $X$ and the quotients $G/H_i$. □

Note that in general a permutation representation is given by the action of a group $G$ on a set $X$ and then extended on $\mathbb{R}^X$ such that, for each $g \in G$ and $x \in X$, an element $e_x$ of the canonical basis of $\mathbb{R}^X$ transforms as $ge_x = e_{gx}$. But the action of $G$ on $X$, i.e., the computation of $gx$, is not explicit and easy to convert in computational terms, while the action of $G$ on one of its quotients $G/H$ provides an algebraic, hence computable, alternative. Furthermore, this notion provides the complete list of possible sets admitting $G$-actions up to isomorphism, which is in bijection with the set of all the quotients of $G$.

Let us now discuss how representation spaces of permutations equivariant networks on sets [124, 95] reduce to a simple instance of the representation space shown in Corollary 3.2.1. Indeed, the symmetric group of $n$ elements, $S_n$, is the set of permutations of $[n]$, and then $S_n$-equivariant networks are able to process sets of elements independently of their order. A complete treatment of the representations of $S_n$ can be found in [101]. Now we want to construct group quotients able to define representations for $S_n$-equivariant networks used in practice. Consider $\lambda = (\lambda_1, \dots, \lambda_l)$ to be a partition of $n$, i.e., a decreasingly ordered tuple of positive integers whose sum is $n$. Define $S_\lambda = S_{\lambda_1} \times \cdots \times S_{\lambda_l}$ as the subgroup of $S_n$ where the $i$th factor permutes the elements form $\sum_{j=1}^{i-1} \lambda_j$ to $\sum_{j=1}^{i} \lambda_j$. Now set $\lambda$ to be the partition $(n-1, 1)$, elements of $S_n/S_{(n-1,1)}$ will be represented by the identity element and all the permutations of $[n]$ that send 1 into an element $i$ of $\{2, \dots, n\}$ which we indicate with $[(1i)]$ (See Definition 2.1.5 for more information about quotients of groups). The action of $\sigma \in S_n$ on $[(1i)] \in S_n/S_{(n-1,1)}$ will be $[(1\sigma(i))]$ where we can identify $[(11)]$ with the identity element. The bijection $[(1i)] \mapsto i$ from $S_n/S_{(n-1,1)}$ to $[n]$ is compatible with the action of $S_n$ on those sets, hence we have an equivariant isomorphism between $\mathbb{R}^{S_n/S_{(n-1,1)}}$ and $\mathbb{R}^n$ which is the standard representation space for permutations equivariant networks on sets [124, 95]. Due to Corollary 3.2.1, $\mathbb{R}^{S_n/S_{(n-1,1)}}$ is one of the admissible representation spaces for $S_n$ on the family of non-odd non-affine continuous activations. In a similar way, in Section 3.4.2, we will see how to obtain representation spaces of IGNs and we will highlight that many other admissible representation space could be possible. It is relevant how it is possible to obtain efficient models through this procedure and which however offers the possibility of building models starting from the simple knowledge of the group of symmetries of the input.

## 3.3 Disentangled Representations

Disentangled representations [27] can be described as follows. An irreducible representation of a $G$-representation $V_i$ is a minimal non-trivial $G$-invariant subspace, and each representation $V_i$ can be decomposed into a direct sum of irreducible spaces (Definition 2.3.8 and Theorem 2.3.9). Composing irreducible representations with each other allows us to construct arbitrary representation spaces and control the number of parameters of each layer at will thanks to Schur's Lemma (Lemma 2.3.10). The easiest way to compose these representations with each other is by doing a direct sum. If then, to define activations, we

choose a basis whose vectors are contained in a single irreducible component we say that such a space is disentangled [27]. As a consequence of Theorem 3.1.3, we obtain the following characterization of disentangle representations.

**Corollary 3.3.1.** *For finite groups and activation functions $f : V \to V$ induced by non-odd non-affine continuous functions, the representation is disentangled if and only if the representation $V$ is a sum of trivial representations.*

*Proof.* The direct sum of trivial representations is clearly disentangled and admissible for Theorem 3.1.3. To prove the inverse implication, thanks to Corollary 3.2.1, we can consider $V$ to be a permutation representation. By disentanglement, we can suppose $V$ to be irreducible and with basis $\mathcal{B} = \{v_1, \dots, v_m\}$ defining $\tilde{f}$. Given $v \in V$, the subspace Span $\left\{\sum_{g\in G} gv\right\}$ is trivial, non-zero because $V$ is a permutations representation, and $G$-invariant. Hence, $V =$ Span $\left\{\sum_{g\in G} gv\right\}$ is trivial. □

Even if composability and control of the number of parameters are particularly good properties of disentangled network, being able to admit only trivial irreducible representations is not achievable in most cases of practical use. Let us consider again the example of representation spaces of permutation equivariant networks on sets [124, 95]. The irreducible decomposition of the standard action of $S_n$ on $\mathbb{R}^n$ is the direct sum of a trivial component and a $(n-1)$-dimensional one. Hence, a linear layer between this input space and a disentangled admissible representation space will send the input contained in the $(n-1)$-dimensional component to 0 due to Schur's Lemma (Lemma 2.3.10). For large $n$ as customary, this eliminates the entire information inside the input in the forward pass of the first layer of the network.

## 3.4 Practical Settings

In this section, we aim at understanding how Theorem 3.1.4 affects and limits the design choices of networks in a relevant practical scenario such as geometric IGNs [54], a broad family of models particularly proficient in processing geometric data such as point-clouds or meshes. We achieve this objective in Section 3.4.3, but beforehand, we need to decompose our task into two more manageable subproblems. Specifically, in Section 3.4.1, we investigate rotation-equivariant networks, a case of networks of that will be necessary to understand the following analysis. In Section 3.4.2, we delve into higher-order IGN, widely employed in practical applications. Finally, in Section 3.4.3, we merge the preceding results to analyze geometric IGN comprehensively.

### 3.4.1 Connected and Compact Groups

Rotation-equivariant neural networks [34] have the ability to process geometrical data tracking orientation and pose. We now discuss how our results affects the

design of this kind of networks. The group of rotations around the origin of $\mathbb{R}^3$ is denoted as SO(3), it can be described as the group of real orthogonal $3 \times 3$ matrices with positive determinant. More generally, let $G = \mathrm{SO}(n)$ be the group of real orthogonal $n \times n$ matrices with positive determinant. It acts on $\mathbb{R}^n$ by left multiplication. Note that $\mathrm{SO}(n)$ is a connected and compact topological group and the presented representation is irreducible and non-trivial [39]. The following Corollary 3.4.1 implies that rotation-equivariant networks are invariant.

**Corollary 3.4.1.** *Let $G$ be compact group and let $G_0$ be the connected component containing the identity element. An admissible $G$-representation for non-affine activation functions is $G_0$-invariant. In particular, if $G$ is connected, an admissible $G$-representation is trivial.*

*Proof.* First, suppose $G$ to be connected. The image of a continuous representation of a compact and connected topological group is a compact and connected matrix group. The only possible representations described by Theorem 3.1.4 are permutation representations and signed permutation representations whose images are finite groups whose only compact and connected subgroup is the one containing only the identity. Hence, the original representation is trivial. If $G$ is not connected, a $G$-representations induced a $G_0$-representations which will be $G_0$-invariant. □

This means that in general it will be possible to create neural networks capable of performing invariant tasks such as classification but not more general equivariant tasks such as segmentation or detection [95]. Further, as SO(3) is a subgroup of Euclidean transformations of $\mathbb{R}^3$, this phenomenon afflicts networks equivariant with respect to general Euclidean transformations, i.e., they will be invariant to the rotational part of Euclidean transformations although possibly sensitive to reflections and translations.

### 3.4.2 On Invariant Graph Networks

Let us now go back at the example proposed in Section 3.2 and generalize it to high-order structures to obtain IGNs. Introduced by [78], they are a class of neural networks equivariant with respect to the symmetric group and their expressiveness is intimately related to graph neural networks [77, 41]. They are permutation equivariant models taking as input a relational structure such as a set, a graph, or an higher-order structure such as simplicial complexes in form of a tensor of the corresponding order. For example, a directed graph with $n$ nodes can be seen as a tensor in $\mathbb{R}^n \otimes \mathbb{R}^n$. Elements of this space can be represented as linear combinations of $e_i \otimes e_j$ for $i, j \in [n]$ and each $\sigma \in S_n$ acts on them permuting their indices simultaneously, $\sigma(e_i \otimes e_j) = e_{\sigma(i)} \otimes e_{\sigma(j)}$. More in general, a $k$-ary relational structure can be represented as a $k$-order tensor in $(\mathbb{R}^n)^{\otimes k}$ with $S_n$ simultaneously acting on each $\mathbb{R}^n$ component. Regardless of the order of the input tensor, intermediate representation spaces may be the sum of tensors of arbitrary order, hence a linear layer of an IGN will be the

direct sum of linear equivariant maps between spaces $(\mathbb{R}^n)^{\otimes k}$ and $(\mathbb{R}^n)^{\otimes h}$ for arbitrary $k$ and $h$ and they admit point-wise activations, hence, by Corollary 3.2.1, they should be able to be represented as $\mathbb{R}^{\prod_i S_n/H_i}$ for some subgroups $H_i < S_n$. In Section 3.2 we have seen how $\mathbb{R}^n \cong \mathbb{R}^{S_n/S_{(n-1,1)}}$. Following [12], we get $(\mathbb{R}^n)^{\otimes k} = \oplus_{t=0}^{k} a_t \, \mathbb{R}^{S_n/S_{(n-t,1^t)}}$ where $a_t$'s are positive integers.

This shows that representation spaces of IGNs are direct sum of $\mathbb{R}^{S_n/S_{(n-t,1^t)}}$ for some $t$. But such $(n-t,1^t)$ are only but a fraction of the partitions of $n$, therefore $S_{(n-t,1^t)}$'s are only but a fraction of the subgroups $S_\lambda$ which are only few of the subgroups of $S_n$ as they are not transitive on $[n]$ for $\lambda \neq (n)$, unlike other subgroups such as the alternating group, $A_n$. Indeed, the module $\mathbb{R}^{S_n/A_n}$ will be a two-dimensional representation compatible with point-wise activations. [77] and [43] prove lower and upper bounds on the expressiveness of $k$-order IGNs. In the proofs of these bounds, they implicitly employ the decomposition $(\mathbb{R}^n)^{\otimes k} = \oplus_{t=0}^{k} a_t \, \mathbb{R}^{S_n/S_{(n-t,1^t)}}$ as this equality is strongly related to the decomposition in tensors of different partition type. This makes it natural to ask what the expressiveness of models employing other types of admissible spaces would be.

### 3.4.3 Geometric Graphs and Product Groups

Geometric graphs [54, 34, 9, 40] are utilized as a data structure for modeling systems in computational biology, computational chemistry, and computer graphics. Those are graphs, or higher-order structures, embedded in the Euclidean space and as such may transform according to the symmetries of the ambient space, i.e., isometries of $\mathbb{R}^3$, $\mathrm{E}(3)$. Hence, it becomes relevant to develop neural networks simultaneously equivariant to such ambient symmetries and node permutations, and in algebraic terms this means that we need $\mathrm{E}(3) \times S_n$-equivariant networks able to process elements in $\mathbb{R}^3 \otimes (\mathbb{R}^n)^{\otimes k}$, where the first tensor factor represent the geometrical features and the second the relational structure. We synthesize the results concerning this case in the following corollary of Theorem 3.1.4.

**Corollary 3.4.2.** *Every* $\mathrm{SO}(3) \times S_n$*-equivariant layer defined on* $\mathbb{R}^3 \otimes (\mathbb{R}^n)^{\otimes k}$ *coupled with non-affine activations is null. Hence,* $\mathrm{E}(3) \times S_n$*-equivariant networks are rotation invariant.*

# Chapter 4

# Spaces of Neural Networks

The previous chapter showed that, for compact groups, only limited classes of representations make point-wise activations equivariant. In particular, only permutation representations admit the larger class of activations; for the rest of this manuscript, we therefore focus on this class of representations.

This chapter presents the necessary background on permutation representations and then introduces the following concepts:

- *Layer spaces*, the spaces of affine maps associated with layers of neural networks, together with further prescribed inductive biases, e.g., filter width in the case of convolutional layers.
- *Neural spaces*, the spaces of functions realized by neural networks with a prescribed architecture.
- *Universality classes*, the spaces of functions that can be approximated by neural networks with a prescribed architecture and variable dimensions of invariant hidden features.

## 4.1 Permutation Representations

For completeness, we recall that permutation representations were already introduced in Definition 2.3.3, but we provide here an alternative definition. We will show the equivalence between the two definitions in Remark 4.1.3.

**Definition 4.1.1** (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 4.1.2.** *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. □

*Remark* 4.1.3. Moreover, after choosing an ordering $X = \{x_1, \ldots, x_n\}$ 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 $\mathrm{GL}_n(\mathbb{R})$.

**Proposition 4.1.4.** *Let $X$ and $Y$ be two $G$-sets. We have the two following $G$-representations isomorphisms*

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

*where $X \sqcup Y$ indicate the disjoint union of the sets $X$ and $Y$.*

**Example 4.1.5.** Let $S = [n]$ and let $S_n$ act on $S$ in the standard way and note that $\mathbb{R}^S \cong \mathbb{R}^n$ as representations. From Proposition 4.1.4, we obtain that tensors of order 2 are $\mathbb{R}^n \otimes \mathbb{R}^n = \mathbb{R}^{S\times S} = \mathbb{R}^{n\times n}$. Let $\Delta = \{(i, i) \mid i \in S\}$ and $\overline{\Delta} = \{(i, j) \in S \mid i \neq j\}$, note that $S \times S = \Delta \sqcup \overline{\Delta}$ and that $S_n$ acts transitively on both $\Delta$ and $\overline{\Delta}$. Therefore,

$$\mathbb{R}^n \otimes \mathbb{R}^n \cong \mathbb{R}^{S\times S} \cong \mathbb{R}^{\Delta} \oplus \mathbb{R}^{\overline{\Delta}}.$$

We say that a group $G$ acts transitively on a set $X$ if this action has only one orbit, namely $Gx = X$ for each $x \in X$. If $X = X_1 \sqcup \cdots \sqcup X_n$ is the orbit decomposition of $X$, then $\mathbb{R}^X \cong \mathbb{R}^{X_1} \oplus \cdots \oplus \mathbb{R}^{X_n}$.

As the bias terms of equivariant layers are vectors invariant under the action of permutation representations, it is important to characterize the invariant part of a permutation representation. Before proceeding, it is necessary to state the following result.

**Proposition 4.1.6.** *If $G$ is a finite group acting transitively on a finite set $X$, then there exists a subgroup $H < G$ and a $G$-set bijection between $X$ and $G/H$.*

Proposition 4.1.6 implies that we can restrict our study to representations of the form $\mathbb{R}^{G/H}$ for some subgroup $H$ of $G$. With this result, we can now proceed to prove Proposition 4.1.7. Let $G$ be a group and $X$ be a finite set with action of $G$.

Thanks to Proposition 4.1.2, equivariant affine maps $\mathrm{Aff}_G(V,W)$ can be decomposed into a linear part in $\mathrm{Hom}_G(V,W)$ and a translational part in the invariant subspace $W^G$, the set of $G$-invariant vectors of $W$. In our setting, the symmetry group $G$ is finite, and $W$ is a permutation representation, with its invariant part $W^G$ characterized by the following result.

**Proposition 4.1.7.** *Let $\mathbb{R}^X$ be a permutation representation of $G$ with orbit decomposition $X_1 \sqcup \cdots \sqcup X_n$ (see Definition 2.2.2), let $Y \subseteq X$. Define $\mathbb{1}_Y = \sum_{y\in Y} e_y \in \mathbb{R}^X$. The invariant subspace of*

$$\mathbb{R}^X = \mathbb{R}^{X_1} \oplus \cdots \oplus \mathbb{R}^{X_n},$$

*consisting of vectors fixed by the action of $G$, is generated by the basis*

$$\mathbb{1}_{X_1}, \ldots, \mathbb{1}_{X_n}.$$

*Proof.* The Reynolds operator

$$\mathcal{R}: \begin{array}{c} V \longrightarrow V^G \\ v \mapsto \displaystyle\sum_{g\in G} gv \end{array}$$

projects each $G$-representation $V$ on its invariant subspace $V^G$. In the case $V = \mathbb{R}^{G/H}$, $e_{kH}$ is an element of the canonical base of $\mathbb{R}^{G/H}$,

$$\mathcal{R}(e_{kH}) = \sum_{g\in G} g e_{kH} = \sum_{g\in G} e_{gkH} = |H| \sum_{gH\in G/H} e_{gH} = |H| \mathbb{1}_{G/H}.$$

The final observation follows from Proposition 4.1.4. □

Propositions 4.1.4 and 2.4.4 together imply that characterizing equivariant maps between permutation representations reduces to characterizing equivariant maps between representations induced by transitive actions on finite sets, or equivalently, left cosets by Proposition 4.1.6. We address this in Proposition 4.1.11. To prove this result, we first define the concept of right multiplication in Definition 4.1.8 and then prove Proposition 4.1.9, which characterizes equivariant maps between regular representations.

**Definition 4.1.8.** For each $g \in G$ define the right-multiplication

$$\mathcal{R}_g: \begin{array}{c} \mathbb{R}^G \longrightarrow \mathbb{R}^G \\ e_h \mapsto e_{hg^{-1}}. \end{array}$$

**Proposition 4.1.9.** *Right actions are a basis for the space of equivariant endomorphisms of the regular representation. In other words, $\{\mathcal{R}_g\}_{g\in G}$ is a basis for* $\mathrm{Hom}_G(\mathbb{R}^G, \mathbb{R}^G)$.

*Proof.* Each linear application $\phi \in \mathrm{Hom}(\mathbb{R}^G, \mathbb{R}^G)$ is defined by the values $\phi(e_g)$ for each $g \in G$ by linear extension. If $\phi$ is $G$-equivariant, it is defined just by its value on $e_e$. Indeed, $\phi(e_g) = \phi(ge_e) = g\phi(e_e)$ for each $g \in G$. Note that $\mathcal{R}_g$ is linear as the right action of $g$ on $\mathbb{R}^G$ is linear and $\mathcal{R}_g(e_h) = e_{hg^{-1}} = e_h g^{-1}$. It is also equivariant, indeed $\mathcal{R}_g(hv) = hvg^{-1} = h\,\mathcal{R}_g(v)$ for each $v \in \mathbb{R}^G$ and $h \in G$. Furthermore, $\mathcal{R}_{g^{-1}}(e_e) = e_g$ for each $g \in G$ and therefore they generate $\mathrm{Hom}_G(\mathbb{R}^G, \mathbb{R}^G)$.

Suppose that there exist values $a_g \in \mathbb{R}$ for each $g \in G$ such that

$$\sum_{g\in G} a_g\,\mathcal{R}_g = 0,$$

then

$$0 = \sum_{g\in G} a_g\,\mathcal{R}_g(e_e) = \sum_{g\in G} a_g e_{g^{-1}}.$$

Since elements $e_g$ are linearly independent, $a_g = 0$ for each $g \in G$. Hence, $\mathcal{R}_g$ are linearly independent and form a basis for $\mathrm{Hom}_G(\mathbb{R}^G, \mathbb{R}^G)$. □

Now we would like to have a result similar to Proposition 4.1.9 but for morphisms between $G/K$ and $G/H$. To do this we need to define the following injection

$$\iota_{G/H} : \begin{array}{c} \mathbb{R}^{G/H} \to \mathbb{R}^G \\ e_{gH} \mapsto \dfrac{1}{|H|}\displaystyle\sum_{h\in H} e_{gh}, \end{array}$$

and projection

$$\pi_{G/H} : \begin{array}{c} \mathbb{R}^G \to \mathbb{R}^{G/H} \\ e_g \mapsto e_{gH}. \end{array}$$

For an arbitrary representation $V$, we define two surjective maps

$$\iota^*_{G/K} : \begin{array}{c} \mathrm{Hom}_G(\mathbb{R}^G, V) \twoheadrightarrow \mathrm{Hom}_G(\mathbb{R}^{G/K}, V) \\ \phi \mapsto \phi \circ \iota_{G/K}, \end{array}$$

and

$$\pi_{G/H*} : \begin{array}{c} \mathrm{Hom}_G(V, \mathbb{R}^G) \twoheadrightarrow \mathrm{Hom}_G(V, \mathbb{R}^{G/H}) \\ \phi \mapsto \pi_{G/H} \circ \phi. \end{array}$$

We can now generalize the concept of right multiplication to the general case of transitive actions, a concept necessary for stating Proposition 4.1.11, which we will then prove by following the approach used in the proof of Proposition 4.1.9.

**Definition 4.1.10.** Define $\mathcal{R}_{HgK} = \alpha\,\mathcal{R}_g$ where the map $\alpha = \pi_{G/H*} \circ \iota^*_{G/K}$ is defined from $\mathrm{Hom}_G(\mathbb{R}^G, \mathbb{R}^G)$ to $\mathrm{Hom}_G(\mathbb{R}^{G/K}, \mathbb{R}^{G/H})$.

**Proposition 4.1.11.** *The map $\mathcal{R}_{HgK}$ is well-defined and the set $\{\mathcal{R}_{HgK}\}_{HgK\in H\backslash G/K}$ is a basis for $\mathrm{Hom}_G(\mathbb{R}^{G/K},\mathbb{R}^{G/H})$. Finally,*

$$\Big(\mathcal{R}_{HgK}(e_{kK})\Big)_{sH}=\begin{cases}\frac{1}{|K|} & \text{if } sH\subseteq kKg^{-1}H,\\ 0 & \text{otherwise.}\end{cases}$$

*Proof.* To prove that $\mathcal{R}_{HgK}$ is well-defined, we need to prove that $\alpha\,\mathcal{R}_g=\alpha\,\mathcal{R}_{hgk}$ for each $h\in H$ and $k\in K$. Indeed,

$$\pi_{G/H}\,\mathcal{R}_{hgk}\,\iota_{G/K}(e_{sK})=\frac{1}{|K|}\sum_{t\in K}e_{stk^{-1}g^{-1}h^{-1}H}=$$
$$\frac{1}{|K|}\sum_{t\in K}e_{stk^{-1}g^{-1}h^{-1}H}=\frac{1}{|K|}\sum_{t\in K}e_{stg^{-1}H}=\pi_{G/H}\,\mathcal{R}_g\,\iota_{G/K}(e_{sK}),$$

where the penultimate equality is true because $h^{-1}H=H$ and variable change $t\mapsto tk^{-1}$ in the sum.

By Proposition 4.1.9 the set $\{\mathcal{R}_g\}_{g\in G}$ is a basis for $\mathrm{Hom}_G(\mathbb{R}^G,\mathbb{R}^G)$. By the previous observation we have shown that the image of $\{\mathcal{R}_g\}_{g\in G}$ under $\alpha$ is $\{\mathcal{R}_{HgK}\}_{HgK\in H\backslash G/K}$. As $\alpha$ is a surjection, $\{\mathcal{R}_{HgK}\}_{HgK\in H\backslash G/K}$ generates $\mathrm{Hom}_G(\mathbb{R}^{G/K},\mathbb{R}^{G/H})$.

Proving linear independence is similar to the proof of linear independence in Proposition 4.1.9. Indeed, let $a_{HgK}\in\mathbb{R}$ for each $HgK\in H\backslash G/K$ such that

$$\sum_{HgK\in H\backslash G/K}a_{HgK}\,\mathcal{R}_{HgK}=0.$$

Hence,

$$0=\sum_{HgK\in H\backslash G/K}a_{HgK}\,\mathcal{R}_{HgK}(e_K)=\sum_{HgK\in H\backslash G/K}\frac{1}{|K|}a_{HgK}\sum_{t\in K}e_{tg^{-1}H}.$$

Note that sets $\{tg^{-1}H\}_{t\in K}$ are pairwise disjoint with $g$ varying between representatives of $HgK$. This means that the respective vectors $\sum_{t\in K}e_{tg^{-1}H}$ are linearly independent, hence each $a_{HgK}=0$. This proves that the maps $\mathcal{R}_{HgK}$ are linearly independent. Finally, observing that

$$\mathcal{R}_{HgK}(e_{kK})=\frac{1}{|K|}\sum_{t\in K}e_{ktg^{-1}H},$$

it is clear that

$$\Big(\mathcal{R}_{HgK}(e_{kK})\Big)_{sH}=\begin{cases}\frac{1}{|K|} & \text{if } sH\subseteq kKg^{-1}H,\\ 0 & \text{otherwise.}\end{cases}$$

□

*Remark* 4.1.12. In our case of interest, in which $G$ is a finite group, the map

$$v \mapsto v \cdot w$$

is equivalent at convolving $v$ by $w$. Proposition 4.1.11 is just a restatement and integration of Theorem 1 in [64] in the restricted case of homogeneous spaces of finite groups.

## 4.2 Equivariant Neural Networks

Although we already presented the definition of equivariant neural networks in (2.2), we now place it in a broader framework of spaces of neural networks, in a way that allows us to discuss architecture while decoupling it from the choice of parameterization.

We now recall the definition of pointwise-activations.

**Definition 4.2.1** (Point-wise Activation)**.** Let $\mathbb{R}^X$ be a permutation representation of a group $G$, and let $\sigma : \mathbb{R} \to \mathbb{R}$. We define the *point-wise activation* induced by $\sigma$ as the function $\tilde{\sigma} : \mathbb{R}^X \to \mathbb{R}^X$ such that $\tilde{\sigma}(\sum_{x\in X} \alpha_x e_x) = \sum_{x\in X} \sigma(\alpha_x) e_x$. We will often abuse notation and refer to $\sigma$ as the activation function as well.

As a first step toward defining spaces of neural networks, we introduce layer spaces.

**Definition 4.2.2** (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 $M \subseteq \mathrm{Aff}_G(V, \mathbb{R}^X)$. We distinguish the following cases:

- We say that $M$ has *complete bias* if, via $\theta$,

  $$\lambda(M) \oplus \tau(M) \cong M \quad \text{and} \quad \tau(M) = \langle \mathbb{1}_P \rangle_{P\in\mathcal{P}}$$

  for some partition $\mathcal{P}$ of $X$. In this case, we say that the bias of $M$ is *subordinate* to $\mathcal{P}$.

- Otherwise, we say that $M$ has *incomplete bias*.

- We say that $M$ has *full bias* if it has complete bias subordinate to the orbit partition of $X$.

- We say that $M$ has *null bias* if $\tau(M) = 0$. Equivalently, $M$ is contained in $\mathrm{Hom}_G(V, \mathbb{R}^X)$ and $M = \lambda(M)$.

*Remark* 4.2.3. A layer space $M$ with complete bias is of the form

$$M = \left\{ v \mapsto \sum_{i=1}^{k} x_i \phi^i(v) + \sum_{P\in\mathcal{P}} y_P \mathbb{1}_P \;\middle|\; x_1, \ldots, x_k \in \mathbb{R},\ y_P \in \mathbb{R} \text{ for all } P \in \mathcal{P} \right\},$$

where $\phi^1, \dots, \phi^k$ generate a subspace of $\mathrm{Hom}_G(V, \mathbb{R}^X)$, and $\mathcal{P}$ is a partition of $X$ that may either combine several orbits from the orbit partition $X = X_1 \sqcup \cdots \sqcup X_n$ into larger subsets, or coincide with the orbit partition itself. Moreover,

$$\mathbb{1}_{X_j} := \sum_{x \in X_j} e_x$$

for $j = 1, \dots, n$, with $\{e_x\}_{x \in X}$ the canonical basis of $\mathbb{R}^X$.

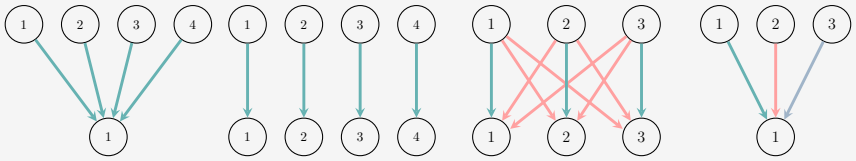
In the following, we will often abuse notation and consider layer spaces to have complete bias, unless specified otherwise.

Note that each subspace $M$ of $\mathrm{Aff}_G(V, \mathbb{R}^X)$ such that $M = \lambda(M) \oplus \tau(M)$ and $\tau(M) = \tau(\mathrm{Aff}_G(V, \mathbb{R}^X))$ has complete bias. Indeed, in this case we have $\mathcal{P} = \{X_1, \dots, X_n\}$, the $G$-orbit decomposition of $X$ and

$$\tau(M) = \langle \mathbb{1}_{X_i} \rangle_{X_i \in \mathcal{P}}$$

as proven in Proposition 4.1.7.

With this further characterization, we present detailed examples of equivariant affine maps relevant to machine learning applications, showing how common layers can be expressed within this formalism.




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

Figure 4.1: Colored bipartite graphs illustrating the layer spaces used in Examples 4.2.6 and 4.2.7, along with the graph corresponding to a standard fully connected layer, following the convention of [99] where arrows of the same color denote identical weights applied to different input values.

**Example 4.2.4** (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

$$L := \mathrm{Aff}(\mathbb{R}, \mathbb{R}),$$

whose relevance will become clearer later, for instance, in relation to (4.3).

**Example 4.2.5** (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

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

**Example 4.2.6** (PointNet Layers)**.** We focus on the layer spaces used in the sum-pooling variant of PointNet architectures [95], 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. [124] 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

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

where $\mathbb{1} = \mathbb{1}_{[n]} = [1, \ldots, 1]^\top$. Figure 4.1b 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 Example 4.2.5. Figure 4.1a shows the colored bipartite graph corresponding to the layer space $\mathrm{Aff}_{S_n}(\mathbb{R}^n, \mathbb{R})$.

**Example 4.2.7** (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 $\mathbb{R}^n$:

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

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:

$$C^k = \left\{ v \mapsto \sum_{i=1}^{k} x_i A(e_i)v + y\mathbb{1}_{[n]} \;\middle|\; x_1, \ldots, x_k, y \in \mathbb{R} \right\}. \tag{4.1}$$

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 $1 \leq k_1, \ldots, k_d \leq n$. One example of particular importance in the following chapters is the case $k = 1$, corresponding to convolutional filters of width 1. This layer space also generalizes to a subset of $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X)$ for an arbitrary permutation representation $\mathbb{R}^X$ with orbit decomposition $X = X_1 \sqcup \cdots \sqcup X_\ell$, in which case we denote it by

$$C := \left\{ v \mapsto \lambda v + \sum_{i=1}^{\ell} \mu_i \mathbb{1}_{X_i} \;\middle|\; \lambda \in \mathbb{R},\ \mu_1, \ldots, \mu_\ell \in \mathbb{R} \right\}. \tag{4.2}$$

The four examples defined above, namely,

- The space $L$ of arbitrary univariate affine maps,
- The space $I$ of invariant affine layers,
- The space $P$ of permutation-equivariant affine maps,
- The space $C$ of convolutional filters of width 1,

will constitute the main prototype examples of layers employed throughout the rest of the manuscript.

We now define the function space associated with a fixed neural architecture, which we call a *neural space*, and which is also referred to in the literature as a *neuromanifold* [22].

**Definition 4.2.8** (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 $\sigma$ is defined by

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

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 $\sigma$.

**Example 4.2.9.** Using the examples of layer spaces introduced above, we can define multiple architectures:

- Equivariant PointNets: $\mathcal{N}_\sigma(P, \dots, P)$,
- Invariant PointNets: $\mathcal{N}_\sigma(P, \dots, P, I)$,
- Convolutional Neural Networks: $\mathcal{N}_\sigma(C^{k_1}, \dots, C^{k_d})$.

## 4.3 Special Constructions

In this section we introduce constructions to obtain new neural spaces by combining and transforming other neural spaces. These constructions will be employed in proofs and serve as tools both to decompose complex architectures into an interaction of simpler spaces, for instance via composition in Chapter 7, and, conversely, to reduce a complex interaction between spaces to a single space, for instance via twin spaces in Chapter 5 or sum spaces in Chapter 6.

### 4.3.1 Composition of Neural Networks

We start by studying compositions of neural spaces with compatible output and input spaces.

**Definition 4.3.1** (Composition)**.** Let $X$ and $Y$ be topological spaces, given two families of functions $\mathcal{F}$ in $\mathcal{C}(X, Y)$ and $\mathcal{G}$ in $\mathcal{C}(Y, Z)$ we define their composition as

$$\mathcal{F} \hat{\circ} \mathcal{G} = \{f \circ g \mid f \in \mathcal{F}, g \in \mathcal{G}\}.$$

Note that the concatenation of two layer spaces is not necessarily a layer space but there is a useful exception:

**Proposition 4.3.2.** *Let $C$ be the layer space of associated to width-1 convolutional filters in $\mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X)$ and $M_1, \dots, M_d$ layer spaces with $M_i$ with full bias and contained in $\mathrm{Aff}_G(\mathbb{R}^X, W)$ then*

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

*Proof.* The proof follows directly from

$$M_i \hat{\circ} C = M_i.$$

Indeed,

$$\begin{aligned} M_i \hat{\circ} C &= \big\{x \mapsto \tau_w\phi\big(\lambda x + v\big) \mid \tau_w\phi \in M_i,\ \lambda \in \mathbb{R},\ v \in (\mathbb{R}^X)^G\big\} \\ &= \big\{x \mapsto \tau_w\big([\lambda\, \phi](x)\big) + \phi(v) \mid \tau_w\phi \in M_i,\ \lambda \in \mathbb{R},\ v \in (\mathbb{R}^X)^G\big\} \\ &= \big\{x \mapsto \tau_{w+\phi(v)}\phi(x) \mid \tau_w\phi \in M_i,\ v \in (\mathbb{R}^X)^G\big\} \\ &= M_i. \end{aligned}$$

□

### 4.3.2 Twin Neural Networks

Here we introduce the concept of twin neural networks and their associated neural spaces. They will be an important tool in Chapter 5, where we will reduce the separation constraint of a neural space to the zero locus of an associated twin space, see Proposition 5.2.1.

**Definition 4.3.3** (Parallelization, Split and Twins)**.** Let $X$ and $Y$ be topological spaces. Define the parallelization map

$$\begin{array}{rccc} \Pi: & \mathcal{C}(X,Y) & \to & \mathcal{C}(X\times X, Y\times Y) \\ & f & \mapsto & \big[(x,y)\mapsto (f(x),f(y))\big]. \end{array}$$

Define the split map

$$\begin{array}{rccc} \Psi: & \mathcal{C}(X,Y) & \to & \mathcal{C}(X, Y\times Y) \\ & f & \mapsto & \big[(x,y)\mapsto (f(x),f(y))\big]. \end{array}$$

Let $V$ be a vector space and define the twin map

$$\begin{array}{rccc} \mathrm{T}: & \mathcal{C}(X,V) & \to & \mathcal{C}(X\times X, V) \\ & f & \mapsto & \big[(x,y)\mapsto f(x)-f(y)\big]. \end{array}$$

We are interested in characterizing the images of these maps when restricted to neural networks. The following definition is useful to better describe the structure of the bias terms of layer spaces transformed by these maps.

**Definition 4.3.4.** Let $X$ be a finite set, we define the *duplicate set* of $X$ as the set $X \sqcup X'$ where $X'$ is a disjoint copy of $X$. Let $\mathcal{P}$ be a partition of $X$, we define the *duplicate partition* of $\mathcal{P}$ as the partition $\mathcal{P}'$ of the duplicate of $X$ such that $\mathcal{P}' = \{Y \sqcup Y' \mid Y \in \mathcal{P}\}$. For each $y \in Y$, we will usually indicate the respective element in $Y'$ as $y'$, although when it will be clear from the context we may abuse notation and call both $y$.

With these definitions in place, restricting $\Pi$, $\Psi$, and T to neural networks yields the following properties.

**Proposition 4.3.5.** *For layer spaces $M_1, \dots, M_d$, we have*

$$\Pi\big(\mathcal{N}_\sigma(M_1,\dots,M_d)\big) = \mathcal{N}_\sigma\big(\Pi(M_1),\dots,\Pi(M_d)\big),$$

$$\Psi\big(\mathcal{N}_\sigma(M_1,\dots,M_d)\big) = \mathcal{N}_\sigma\big(\Psi(M_1),\Pi(M_2),\dots,\Pi(M_d)\big),$$

*and*

$$\mathrm{T}\big(\mathcal{N}_\sigma(M_1,\dots,M_d)\big) = \mathcal{N}_\sigma\big(\Psi(M_1),\Pi(M_2),\dots,\Pi(M_{d-1}),\mathrm{T}(M_d)\big).$$

*Moreover, if $M$ is a layer space and has complete bias subordinate to a partition $\mathcal{P}$, then $\Pi(M)$ and $\Psi(M)$ have complete bias subordinate to the duplicate partition of $\mathcal{P}$. In particular, if $M$ has complete bias, then $\Pi(M)$ and $\Psi(M)$ also have complete bias. However, $M$ having full bias does not imply that $\Pi(M)$ and $\Psi(M)$ have full bias. Note that for an arbitrary layer space $M$, the associated twin layer space $\mathrm{T}(M)$ has null bias. Finally,*

$$\dim M = \dim \Pi(M) = \dim \Psi(M).$$

### 4.3.3 Sums of Neural Networks

Here we introduce the concept of neural spaces associated to sums of neural networks. They will be an important tool in what follows, culminating in Proposition 4.3.10, which shows that relevant families of neural networks are vector spaces, enabling us to employ functional analysis to study their approximation capabilities.

**Definition 4.3.6** (Diagonalization, Fork and Sum)**.** Let $X$ and $Y$ be topological spaces. Define the diagonalization map

$$\Delta : \begin{array}{rcl} \mathcal{C}(X,Y) \times \mathcal{C}(X,Y) & \to & \mathcal{C}(X, Y \times Y) \\ (f,g) & \mapsto & \big[x \mapsto (f(x), g(x))\big]. \end{array}$$

Define the fork map

$$\Phi : \begin{array}{rcl} \mathcal{C}(X,Y) \times \mathcal{C}(X,Y) & \to & \mathcal{C}(X, Y \times Y) \\ (f,g) & \mapsto & \big[x \mapsto (f(x), g(x))\big]. \end{array}$$

Let $V$ be a vector space and define the sum map

$$\Sigma : \begin{array}{rcl} \mathcal{C}(X,V) \times \mathcal{C}(X,V) & \to & \mathcal{C}(X, V) \\ (f,g) & \mapsto & \big[x \mapsto f(x) + g(x)\big]. \end{array}$$

For a family of functions $\mathcal{F}$, we will write $\Delta(\mathcal{F}) := \Delta(\mathcal{F}, \mathcal{F})$. We will also adopt a similar notation for $\Phi$ and $\Sigma$.

Note that the definitions of $\Delta$, $\Phi$, and $\Sigma$ are similar to those of $\Pi$, $\Psi$, and T, respectively, but with significant differences. Indeed, the maps $\Pi$, $\Psi$, and T associate a single map to a function of two variables, whereas $\Delta$, $\Phi$, and $\Sigma$ associate a pair of maps to a function of a single variable. Restricting $\Delta, \Phi$ and $\Sigma$ to neural networks yields the following properties.

**Proposition 4.3.7.** *For layer spaces $M_1, \ldots, M_d$, we have*

$$\Delta\big(\mathcal{N}_\sigma(M_1, \ldots, M_d)\big) = \mathcal{N}_\sigma\big(\Delta(M_1), \ldots, \Delta(M_d)\big),$$

$$\Phi\big(\mathcal{N}_\sigma(M_1, \ldots, M_d)\big) = \mathcal{N}_\sigma\big(\Phi(M_1), \Delta(M_2), \ldots, \Delta(M_d)\big),$$

*and*

$$\begin{aligned} &\Sigma\big(\mathcal{N}_\sigma(M_1, \ldots, M_d)\big) \\ &= \mathcal{N}_\sigma\big(\Phi(M_1),\ \Delta(M_2),\ \ldots,\ \Delta(M_{d-1}),\ \Sigma(M_d)\big). \end{aligned}$$

*Note that if $M$ is a layer space with full, complete, or null bias, then $\Delta(M)$, $\Phi(M)$, and $\Sigma(M)$ are also layer spaces with full, complete, or null bias, respectively.*

### 4.3.4 Block Spaces

The final construction we introduce is needed as a surrogate for width, or the dimension of invariant hidden features, in contexts where such quantities are not defined. However, these quantities are necessary to discuss universality of neural networks when those notions are otherwise undefined.

**Definition 4.3.8** (Block Spaces)**.** Let $W$ be a vector space and let $\mathcal{F} \subseteq \mathcal{C}(X, W)$ be a family of functions. For each $h, k \in \mathbb{N}$, we define the associated $h \times k$ block space as

$$\mathcal{F}^{k\times h} := \left\{ \begin{array}{c} (x_1, \dots, x_k) \mapsto \left(\sum_{j=1}^k f_{1,j}(x_j), \dots, \sum_{j=1}^k f_{h,j}(x_j)\right) \\ f_{ij} \in \mathcal{F}, \quad i = 1, \dots, h, \quad j = 1, \dots, k \end{array} \right\}. \tag{4.3}$$

Note that if $\mathcal{F}$ is a layer space $M \subseteq \mathrm{Aff}_G(V, W)$ for some $G$-representations $V$ and $W$, then

$$M^{k\times h} \subseteq \mathrm{Aff}_G(V \otimes \mathbb{R}^k, W \otimes \mathbb{R}^h).$$

Here $k$ can be interpreted as a surrogate width, or equivalently as the dimension of the hidden invariant features of the input space, while $h$ plays the analogous role for the output space. In particular, if $f \in M^{k\times h}$, then $f$ can be written as in (4.3) as

$$f : (x_1, \dots, x_k) \mapsto \Big(\sum_{j=1}^k f_{1,j}(x_j), \dots, \sum_{j=1}^k f_{h,j}(x_j)\Big)$$

for some $f_{ij} \in M$, which can be written as

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

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 $f$ can be written respectively as

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

In other words, the linear part of $M^{k\times h}$ is an $h \times k$ block matrix whose blocks belong to the linear part of $M$.

**Example 4.3.9.** 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.

It is also interesting to note that

$$\Pi(M) \subseteq \Phi(M) = M^{1\times 2}, \qquad \Delta(M) \subseteq M^{2\times 2}, \qquad \mathrm{T}(M),\ \Sigma(M) \subseteq M^{2\times 1}.$$

In particular, Proposition 4.3.7 implies the following.

**Proposition 4.3.10.** *For layer spaces $M_1, \ldots, M_d$ and each*

$$f, g \in \mathcal{N}_\sigma(M_1, \ldots, M_d),$$

*we have*

$$f + g \in \mathcal{N}_\sigma\big(M_1^{1\times 2}, M_2^{2\times 2}, \ldots, M_{d-1}^{2\times 2}, M_d^{2\times 1}\big).$$

*In particular, the space of all neural networks with variable width in each layer, namely,*

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

*is a vector space.*

## 4.4 Universality Classes

Proposition 4.3.10 suggests the possibility of studying the space of all neural networks with a fixed architecture and variable surrogate width in each layer. The main question of this thesis is to understand which functions can be approximated by this class.

We start by recasting known results in the language of neural spaces, in order to better introduce the concept of universality class. The first example is shallow multilayer perceptrons. Observe that the class of shallow neural networks with variable width can be written as $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(L^{1\times h}, L^{h\times 1})$. We denote by $\mathcal{U}_\sigma(L, L)$ the associated universality class, namely the set of continuous functions approximable by such networks. Formally, $\mathcal{U}_\sigma(L, L)$ is defined as the closure of this union in $\mathcal{C}(\mathbb{R})$, equipped with the topology of uniform convergence on compact sets.

**Theorem 4.4.1** ([90])**.** *The universality class for shallow neural networks, $\mathcal{U}_\sigma(L, L)$, coincides with $\mathcal{C}(\mathbb{R})$ if and only if the activation function $\sigma$ is not polynomial.*

An analogous result in the equivariant setting was established by [98] for neural networks with input and output representations $V$ and $W$, respectively, and regular hidden representations of the form $\mathbb{R}^G$. Define the layer spaces

$$M = \mathrm{Aff}_G(V, \mathbb{R}^G) \quad \text{and} \quad N = \mathrm{Aff}_G(\mathbb{R}^G, W).$$

We define the universality class $\mathcal{U}_\sigma(M, N)$ as the set of functions in $\mathcal{C}(V, W)$ that can be approximated by elements of $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(M, N)$. 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 [98] can then be stated as follows.

**Theorem 4.4.2** ([98])**.** *The universality class $\mathcal{U}_\sigma(M, N)$ coincides with the set $\mathcal{C}_G(V, W)$ of all continuous $G$-equivariant functions from $V$ to $W$, if and only if the activation function $\sigma$ is not a polynomial.*

The following definition of universality classes encompasses the notions introduced in Theorem 4.4.1 and Theorem 4.4.2, while being general enough to cover a wider range of architectures, such as PointNets and CNNs with variable filter size.

**Definition 4.4.3** (Universality Classes)**.** The *universality class* 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.

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.

Thanks to Proposiion 4.3.10, we obtain the following.

**Proposition 4.4.4.** *Let $M_1, \dots, M_d$ be layer spaces such that $M_i$ is contained in $\mathrm{Aff}(V_{i-1}, V_i)$ for each $i = 1, \dots, d$. The associated universality class*

$$\mathcal{U}_\sigma(M_1, \dots, M_d) \subseteq \mathcal{C}(V_0, V_d)$$

*is a closed subspace.*

The following chapters will be entirely dedicated to characterizing these spaces.

# Chapter 5

# Separation Constraints

The usual notion of universality is not directly applicable to neural networks that incorporates invariances of the data [21], since they necessarily act by identifying pairs of inputs that are equivalent under the given set of transformations. This feature creates a complex interaction between the network's ability to discriminate different input data, and the invariant or equivariant structure that they are trying to preserve. Assessing expressivity thus requires first a fine-grained analysis of the separation power of these families of neural networks, namely their capacity of distinguishing distinct inputs, which is a necessary condition for the universality of the models [54].

In the graph learning community, which is a paramount domain where invariant and equivariant models are studied [78, 94, 14], networks are required to be invariant or equivariant under the group of permutations of the graph's nodes. In this domain, the primary methods for comparing separation power are the Weisfeiler-Leman (WL) isomorphism test [120] and homomorphism counting [73]. Significant attention has been devoted to studying this property for graph learning models such as Graph Neural Networks (GNNs) [105, 46, 61], Invariant Graph Networks (IGNs) [78, 80], and subgraph GNNs [2, 14]. However, the WL test and homomorphism counting, along with their variants, have severe limitations imposed by their combinatorial nature. In particular, recent research [54] has highlighted the necessity of developing expressivity measures applicable to models that process data beyond relational structures, such as geometric graphs.

This chapter contribute to this effort by studying the separation power of a more general class of equivariant neural networks. Specifically, we precisely describe the set of input pairs identified by relevant families of neural networks. In contrast, other approaches, limited to IGNs and graph processing, provide only upper bounds on expressiveness [41] or lower bounds that require networks with large hidden feature widths [77]. Additionally, we show how hyperparameter and architectural choices impact the separation power of equivariant neural network models, both in general settings and in specific cases of practical interest.

To study the separation power of relevant classes of equivariant networks, we show that the set of identified points corresponds to the set of common zeros of a modified set of networks (Section 5.2.1). We characterize the set of input pairs identified by these families of neural networks by introducing an explicit formula which is recursive over the networks depth (Section 5.2.2). This result provides important insights into how different hyperparameters and architectural choices impact the design of practical equivariant neural network models. In particular, we prove that any non-polynomial activation is equivalent in separation power, achieving the maximum separability for networks with a fixed architecture (Section 5.2.3). We show that increasing depth enhances separation up to a certain depth, where separation power stabilizes (Section 5.2.4). Furthermore, we prove that the multiplicity of the blocks in hidden representations or, equivalently, the width of invariant hidden features does not affect the separation power of the networks (Section 5.2.5). We demonstrate that the separation power of different block types forms a hierarchy, corresponding to the partial ordering of sub-groups of the symmetry group with respect to which the model is equivariant (Section 5.2.6). We illustrate how these general results apply to practical models (Section 5.3). Specifically, we strengthen existing results by showing that a much broader class of IGNs matches the separation power of WL (Section 5.3.1). Then, we demonstrate that the separation power of circular CNNs depends on the filter size (Section 5.3.2).

The contributions of this chapter can be summarized as follows:

- We address the separation power of equivariant neural networks by fully characterizing the set of points identified by networks with a fixed architecture (Proposition 5.2.1 and Theorem 5.2.2).
- We prove that any continuous, real, element-wise, non-polynomial activation is equivalent in separation power, achieving the maximum separability for networks with a fixed architecture (Theorem 5.2.5).
- We show that increasing depth enhances separation power up to a specific threshold, beyond which it stabilizes (Theorem 5.2.6).
- We illustrate how block decomposition of layers influences separability (Theorem 5.2.8) and how separation power is independent of invariant hidden features (Remark 5.2.9). Notably, this result implies that any $k$-IGN matches the separation power of $k$-WL, improving upon previous results that required IGNs to have large hidden feature widths [77].
- Finally, we show that the minimal components from this decomposition form a hierarchy in separation power (Theorem 5.2.11).

## 5.1 The Separation Constraint

We begin by introducing the notion of separation for families of real-valued functions. Building on this definition, we then generalize to the general case.

**Real Codomains:** The universality property of neural networks enables them to approximate any continuous function with arbitrary precision, meaning there exists a sequence of networks that converges pointwise to each continuous function. Equivariant neural networks are designed to handle target functions with specific structures, represented by transformations that recognize equivalent inputs. However, this characteristic necessitates a deeper examination of their separation power. The separation power $\rho(\mathcal{N})$ of a subset $\mathcal{N} \subseteq \mathcal{C}(X, Y)$ of continuous functions between topological spaces $X$ and $Y$ is defined as follows.

**Definition 5.1.1.** A function $f : X \to Y$ is said to *separate* two points $\alpha, \beta \in X$ if $f(\alpha) \neq f(\beta)$. A family of functions $\mathcal{N}$ from $X$ to $Y$ *separates* $\alpha, \beta \in X$ if there exists a function $f \in \mathcal{N}$ that separates $\alpha$ and $\beta$. If a function or a family of functions fails to separate two points, we say that it *identifies* them. The set of pairs of points that are identified by $\mathcal{N}$ define on an equivalence relation

$$\rho := \rho(\mathcal{N}) = \{(\alpha, \beta) \in X \times X \mid f(\alpha) = f(\beta) \text{ for each } f \in \mathcal{N}\}. \tag{5.1}$$

When working with spaces of neural space, their separation power transfers to the class of functions they can approximate, as shown by the following fact.

*Fact.* Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions in $\mathcal{N}$ that converges pointwise to $f$. If $\alpha, \beta \in X$ such that $f_n(\alpha) = f_n(\beta)$ for all $n \in \mathbb{N}$, then $f(\alpha) = f(\beta)$.

In particular, $\mathcal{N}_\sigma$ cannot approximate with arbitrary precision functions beyond ones respecting $\rho$, namely,

$$\mathcal{C}_\rho(X, Y) := \{f \in \mathcal{C}(X, Y) \mid f(\alpha) = f(\beta)\ \forall(\alpha, \beta) \in \rho(\mathcal{N})\}.$$

Understanding however if the entire set $\mathcal{C}_\rho(X, Y)$ can be approximated leads to the study of *separation-constrained universality*.

Notably, [79] and [98] illustrate this phenomenon in the context of equivariant neural networks, which are proven to approximate any continuous equivariant function. However, their constructions involve intermediate representations of impractically large dimensions. In contrast, [41] and [77] show that permutation equivariant networks commonly used in practice can approximate continuous permutation equivariant functions whose separation power is equivalent to the WL test.

In this chapter, *we address the problem of characterizing $\rho$ for relevant families of equivariant neural networks*, as it is *necessary*, though not sufficient, to understand separation-constrained universality. Specifically, we focus on how hyperparameter and architecture choices influence separability, as we will discuss in Section 5.1.1.

We can now generalize the notion of separation to the general case.

**Arbitrary Codomain:** 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 7.1. 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 $\pi_x$ 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 5.1.2 shows that the interaction between equivariance and the global separation $\rho$ is non-trivial when projecting functions onto different output entries.

**Proposition 5.1.2.** *Let $\rho = \rho(\mathcal{N})$ be the separation relation of a family of equivariant neural networks $\mathcal{N}$. The restriction of $\pi_x$ 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 $\rho$.*

The proof for Proposition 5.1.2 and of all subsequent results may be found in the Appendix D.4.2.

Proposition 5.1.2 shows that, after projection onto a single output coordinate, the space of equivariant functions with separation $\rho$ is constrained by a stricter relation. This relation combines $\rho$ 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 5.1.3** (Separation for CNNs)**.** Let $C$ be the layer space of convolutional filters with width 1 defined in Example 4.2.7. 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, (4.1) 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 D.4.7) that

$$\mathcal{U}_\sigma^d(C) = \{(x_1, \dots, x_n) \mapsto (f(x_1), \dots, f(x_n)) \mid f \in \mathcal{C}(\mathbb{R})\} \subsetneq \mathcal{C}_{S_n}(\mathbb{R}^n, \mathbb{R}^n). \quad (5.2)$$

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, (5.2) 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 5.1.3 shows that characterizing equivariant universality classes requires a finer notion of separability, which we now define.

**Definition 5.1.4** (Separation — Revised)**.** Let $G$ be a finite group acting on a finite set $X = \{x_1, \dots, x_n\}$, and let $\mathbb{R}^X$ denote the associated permutation representation. Let $V$ be another permutation representation over $G$ and $\mathcal{N}$ a 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 *separation constraint* or *separation relation* of $\mathcal{N}$ as the collection of relations

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

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

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

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

As noted in Section 4.4, the previous notion of separation is a necessary condition for approximation, and now we see that this generalized version is necessary as well but stronger. Note that in certain cases 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

$$\rho_{x_1}(\mathcal{N}) = \cdots = \rho_{x_n}(\mathcal{N}).$$

Yet, Example 5.1.3 shows that separation can, in fact, be strictly stronger than standard separation. Indeed, on the one hand (5.2) gives

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

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 decompose $\mathbb{R}^n = \mathbb{R} \times \mathbb{R}^{n-1}$, 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 $\rho_i$, with $i = 2, \dots, n$. This proves the following proposition and shows that the universality class in Example 5.1.3 can be completely characterized by the revised notion of separation universality.

**Proposition 5.1.5.** *Define $\rho = \rho\big(\mathcal{U}_\sigma^d(C)\big)$. Then, $\mathcal{U}_\sigma^d(C) = \mathcal{C}_\rho(\mathbb{R}^n, \mathbb{R}^n)$.*

### 5.1.1 Elements Affecting the Separation Relation

From a practical viewpoint it is fundamental to understand the hyperparameters and architecture choices that affect the separation or approximation power of families of neural networks. For example, IGN's separation and approximation power are influenced by two hyperparameters $(k, w)$, where $k$ represents the network's relational order and $w$ denotes the width of the readout multi-layer perceptron. Informally, [79] showed that IGN $= \cup_{k,w} k\text{-IGN}_w$ is universal for continuous equivariant functions, while [41] proved that $k$-IGN $= \cup_w k\text{-IGN}_w$ is universal only within the class of equivariant functions in $\mathcal{C}_{k\text{-WL}}(X, Y)$, the set of continuous equivariant functions with the same separation power as $k$-WL. This example highlights that, in a general equivariant setting, two types of hyperparameters and architecture choices may exist:

- Those like $k$, which regulate the separation power and, hence, have a huge impact on approximation power, but also have a significant impact on the required computational resources.

- Hyperparameters like $w$, which do not affect separability but may impact separation-constrained approximation, often with a limited impact on computational resources. In what follows we aim to identify which hyperparameters and architecture choices control separation power, determining which belong to the first category and which may fall into the second.

## 5.2 Characterization Theorem

We begin by describing and formulating the twin network trick in Section 5.2.1, which serves as the primary tool for converting a separation problem into a zero locus problem, to be addressed informally in Section 5.2.2. In the subsequent sections, we will explore the implications of this result and how it can be applied to effectively compare the separation power of different neural spaces.

### 5.2.1 From the Separation to Zero Loci

In this section, we introduce the twin network trick, which transforms a network separation problem into a zero locus problem for neural networks. This allows us to apply the recursive techniques for solving zero locus problems developed in Section 5.2.2. Specifically, a zero locus problem involves identifying all points that are mapped to zero by all networks within a given neural space.

More precisely, the separation equivalence relation in (5.1) can be reformulated as the following zero locus problem: $(\alpha, \beta) \in \rho(\mathcal{N}_\sigma(M_1, \dots, M_d))$ if and only if

$$\eta(\alpha) - \eta(\beta) = 0 \ \ \forall \eta \in \mathcal{N}_\sigma(M_1, \dots, M_d). \tag{5.3}$$

We observe that (5.3) reduces to a zero-locus problem involving *twin networks*, see Definition 4.3.3. The twin network associated with a network $\eta$ is defined as

$$\mathrm{T}(\eta)(\alpha, \beta) = \eta(\alpha) - \eta(\beta).$$

It is itself a neural network with the same depth as $\eta$, but with a different architecture. Recalling Proposition 4.3.5, the space of twin networks associated with a given neural space is again a neural space, namely,

$$\mathrm{T}\big(\mathcal{N}_\sigma(M_1, \dots, M_d)\big) = \mathcal{N}_\sigma\big(\Psi(M_1), \Pi(M_2), \dots, \Pi(M_{d-1}), \mathrm{T}(M_d)\big). \tag{5.4}$$

In summary, (5.3) together with (5.4) can be synthesized into the following proposition, which directly links the separation relation to a zero locus.

**Proposition 5.2.1.** *For a family $\mathcal{F}$ of functions between a set $V$ and a vector space $W$, let*

$$\mathcal{V}(\mathcal{F}) := \{\beta \in V \,|\, \eta(\beta) = 0 \,\forall \eta \in \mathcal{F}\}$$

*be the* zero locus *of $\mathcal{F}$. Then, for any neural space $\mathcal{N}_\sigma(M_1, \dots, M_d)$, we have*

$$\rho\big(\mathcal{N}_\sigma(M_1, \dots, M_d)\big) = \mathcal{V}\big(\mathrm{T}\big(\mathcal{N}_\sigma(M_1, \dots, M_d)\big)\big).$$

Our task, then, is to determine the zero locus corresponding to the neural space of twin networks.

### 5.2.2 Characterization of the Zero Locus

In the previous sections, thanks to Proposition 5.2.1, we have translated the problem of computing the separation relation $\rho(\mathcal{N}_\sigma(M_1, \dots, M_d))$ into the problem of computing the zero locus $\mathcal{V}\big(\mathrm{T}\big(\mathcal{N}_\sigma(M_1, \dots, M_d)\big)\big)$. More generally, this zero locus can be determined using the recursive formula proposed in Theorem 5.2.2.

We begin by recalling and defining the necessary notation to state Theorem 5.2.2. Let $M_i$ be vector subspaces of $\mathrm{Aff}_G(\mathbb{R}^{X_{i-1}}, \mathbb{R}^{X_i})$ for $i = 1, \dots, d$. Recall that $\lambda(M_d)$ denotes the linear part of $M_d$, and let $\phi^{d,1}, \dots, \phi^{d,s_d}$ be linear maps spanning $\lambda(M_d)$, and recall that $\tau(M_d) = \langle \mathbb{1}_P \rangle_{P \in \mathcal{P}}$ for some partition $\mathcal{P}$ of $X_d$. Let $\mathcal{Q}$ be another partition of $X_d$; if $\mathcal{Q}$ is finer than $\mathcal{P}$ we indicate this relationship as $\mathcal{Q} \leq \mathcal{P}$. Furthermore, for each $h = 1, \dots, s_d$ and $k \in X_d$ define the family of partitions of $X_d$

$$\Psi_{h,k} = \left\{ \mathcal{Q} \leq \mathcal{P} \mid \sum_{i \in P} \phi^{d,h}_{ki} = 0, \, \forall P \in \mathcal{Q} \right\}.$$

Let $\pi_i : \mathbb{R}^{X_{d-1}} \to \mathbb{R}$ denote the projection onto the $i$-th component of $\mathbb{R}^{X_{d-1}}$ for each $i$ in $X_{d-1}$. For each $i, j \in X_{d-1}$, define the set

$$(M_{d-1})_{ij} = \{\phi' : x \mapsto \pi_i \phi(x) - \pi_j \phi(x) \,|\, \phi \in M\},$$

which represents scalar-valued layers obtained as the differences between the $i$-th and $j$-th part of the $(d-1)$-th layer.

Proposition 4.3.5 shows that we only need to solve zero locus problems for networks with complete bias in the intermediate layers and null bias in the final layer. We are now able to give the complete and formal statement the characterization theorem.

**Theorem 5.2.2.** *Let $M_1, \dots, M_{d-1}$ be layer spaces with complete bias and let $M_d$ be a layer space with null bias. Let $\phi^{d,1}, \dots, \phi^{d,s_d}$ be a set of generators of $M_d$ in $\mathrm{Aff}_G(\mathbb{R}^{X_d}, \mathbb{R}^{X_{d+1}})$, and let the bias of $M_{d-1}$ be subordinate to the partition $\mathcal{P}$. Furthermore, for each $h = 1, \dots, s_d$ and each $k \in X_{d+1}$ define*

$$\Psi_{h,k} = \left\{ \mathcal{Q} \leq \mathcal{P} \mid \sum_{i \in P} \phi^{d,h}_{ki} = 0,\ \forall P \in \mathcal{Q} \right\}.$$

*If $\sigma$ is a non-polynomial activation function, then we have the following recursive formula with respect to network depth*

$$\mathcal{V}\big(\mathcal{N}_\sigma(M_1, \dots, M_d)\big) = \bigcap_{h,k} \bigcup_{\mathcal{Q} \in \Psi_{h,k}} \bigcap_{\substack{P \in \mathcal{Q} \\ i,j \in P}} \mathcal{V}\big(\mathcal{N}_\sigma(M_1, \dots, M_{d-2}, (M_{d-1})_{ij})\big), \tag{5.5}$$

*where*

$$(M_{d-1})_{ij} = \{\phi' : x \mapsto \pi_i \phi(x) - \pi_j \phi(x) \mid \phi \in \lambda(M_{d-1})\},$$

*and $\pi_i : \mathbb{R}^X \to \mathbb{R}$ is the projection on the $i$-th component of $\mathbb{R}^X$ for each $i$ in $X$, and $\lambda(M_{d-1})$ is the linear part of $M_{d-1}$.*

*Proof.* Denote $\mathcal{F}_d = \{\phi^{d,1}, \dots, \phi^{d,s_d}\}$. We can restrict to compute the smaller space $\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d))$ since, by Lemma D.2.7, we know

$$\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_d)) = \mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d)).$$

Each $d$-layer neural network $\eta^{d,h}$ in $\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d)$ can be written, for each input $\beta$, as

$$\eta^{d,h}(\beta) = \phi^{d,h} \tilde{\sigma}(\eta^{d-1}(\beta) + y) \qquad (\forall h = 1, \dots, s_d) \tag{5.6}$$

where

- The map $\phi^{d,h}$ is the $h$-th element in $\mathcal{F}_d$ and is linear since $M_d$ has null bias.
- The the map $\eta^{d-1}$ is $(d-1)$-layer network belonging to
$$\mathcal{N}_\sigma(M_1, \dots, M_{d-2}, \lambda(M_{d-1})).$$
- The vector $y$ is a bias term in the translational part of $M_{d-1}$, namely the invariant sub-space of $\mathbb{R}^{X_d}$, and has complete bias subordinate to a partition $\mathcal{Q}$. Hence,
$$y = \sum_{P \in \mathcal{Q}} y_P \mathbb{1}_P. \tag{5.7}$$

In a similar fashion, define $\eta^{d-1,t}$ in $\mathcal{N}_\sigma(M_1,\dots,M_{d-1},\lambda(M_{d-2}))$ for each $t = 1,\dots,s_{d-1}$ and some $s_{d-1} \geq 1$. Note that

$$\eta^{d-1} = \sum_{t=1}^{s_{d-1}} x_t \eta^{d-1,t} \tag{5.8}$$

for some $x_1,\dots,x_{s_{d-1}} \in \mathbb{R}$. Therefore, by substituting both (5.7) and (5.8) into (5.6), we get

$$\eta^{d,h}(\beta) = \phi^{d,h}\tilde{\sigma}\left(\sum_{t=1}^{s_{d-1}} x_t \eta^{d-1,t}(\beta) + y\right). \tag{5.9}$$

Recall that $\phi^{d,h}$ is a linear map from $\mathbb{R}^{X_d}$ to $\mathbb{R}^{X_{d+1}}$, defined by the elements $\phi^{d,h}_{ki} = \phi^{d,h}_k(e_i)$ for each input entry $i \in X_d$ and output entry $k \in X_{d+1}$.

With this notation, we can express (5.9) in coordinates as follows

$$\eta^{d,h}_k(\beta) = \sum_{i\in X_d} \phi^{d,h}_{ki}\sigma\left(\sum_{t=1}^{s_{d-1}} x_t \eta^{d-1,t}_i(\beta) + y_i\right). \tag{5.10}$$

For each $i \in X_d$, let $P$ be the unique element in $\mathcal{Q}$ containing $i$. Then $y_i = y_P$, where $y_i$ is the coefficient defined in (5.10), and $y_P$ is the one defined in (5.7).

Hence, we can write (5.10) as follows

$$\eta^{d,h}_k(\beta) = \sum_{\substack{P\in\mathcal{Q}\\ i\in P}} \phi^{d,h}_{ki}\sigma\left(\sum_{t=1}^{s_{d-1}} x_t \eta^{d-1,t}_i(\beta) + y_P\right)$$

for each output entry $k$ in $\mathbb{R}^{X_{d+1}}$.

Thus, an element $\beta$ belongs to $\mathcal{V}(\mathcal{N}_\sigma(M_1,\dots,M_{d-1},\mathcal{F}_d))$ if and only if

$$\sum_{\substack{P\in\mathcal{Q}\\ i\in P}} \phi^{d,h}_{ki}\sigma\left(\sum_{t=1}^{s_{d-1}} x_t \eta^{d-1,t}_i(\beta) + y_P\right) = 0 \tag{5.11}$$

for each $x_t, y_P$, $h$, $k$, and $\eta^{d-1,t}$.

Assuming that $\sigma$ is non-polynomial and setting $a_i = \phi^{d,h}_{ki}$ and $b_i = (\eta^{d-1,t}_i(\beta))_t$, the second part of Theorem C.1.1 implies that $(\eta^{d-1,t}_i(\beta))_{i,t}$ solves (5.11) for specific $h$ and $k$ if and only if

$$(\eta^{d-1,t}_i(\beta))_i \in \bigcup_{\mathcal{Q}\in\Psi_{h,k}} \bigcap_{\substack{P\in\mathcal{Q}\\ i,j\in P}} \left\{(\eta^{d-1,t}_i(\gamma))_i \mid \eta^{d-1,t}_i(\gamma) - \eta^{d-1,t}_j(\gamma) = 0\right\}.$$

Note that $\beta$ satisfies (5.11) for specific $h$ and $k$ if and only if $(\eta^{d-1,t}(\beta))_{i,t}$ satisfies it. Hence,

$$\beta \in \bigcup_{\mathcal{Q}\in\Psi_{h,k}} \bigcap_{\substack{P\in\mathcal{Q}\\ i,j\in P}} \left\{\gamma \mid \eta^{d-1,t}_i(\gamma) - \eta^{d-1,t}_j(\gamma) = 0\ \forall t\right\}. \tag{5.12}$$

By the definition of $(M_{d-1})_{ij}$, we get

$$\mathcal{V}(\mathcal{N}_\sigma(M_1, \ldots, M_{d-2}, (M_{d-1})_{ij})) = \{\beta \mid \eta_i^{d-1,t}(\beta) - \eta_j^{d-1,t}(\beta) = 0\}.$$

Therefore, $\beta$ satisfies (5.11) for specific $h$ and $k$ if and only if

$$\beta \in \bigcup_{\mathcal{Q} \in \Psi_{h,k}} \bigcap_{\substack{P \in \mathcal{Q} \\ i,j \in P}} \mathcal{V}(\mathcal{N}_\sigma(M_1, \ldots, M_{d-2}, (M_{d-1})_{ij})).$$

Since $\beta$ has to satisfy (5.11) for each $h$ and $k$, we finally get

$$\mathcal{V}(\mathcal{N}_\sigma(M_1, \ldots, M_d)) = \bigcap_{h,k} \bigcup_{\mathcal{Q} \in \Psi_{h,k}} \bigcap_{\substack{P \in \mathcal{Q} \\ i,j \in P}} \mathcal{V}(\mathcal{N}_\sigma(M_1, \ldots, M_{d-2}, (M_{d-1})_{ij})).$$

□

*Remark* 5.2.3. Theorem 5.2.2 could actually be stated with different activation functions for each layer, as long as they are all non-polynomial. However, for readability and simplicity, we have presented the results using a single activation function.

*Remark* 5.2.4. Here, we demonstrate that the complete bias assumption is necessary for all non-polynomial activations to achieve maximal separation power. Specifically, let us examine the separation power of the set of shallow neural networks where all representation spaces are one-dimensional and the hidden layer has a null, and therefore incomplete, bias term. The main concern is the separability of opposite inputs $\beta$ and $-\beta$. This reduces to study the separation equation

$$y\sigma(\beta x) = y\sigma(-\beta x)$$

for each $x, y \in \mathbb{R}$. Any even function $\sigma$, including non-polynomial ones, solves this equation but does not achieve maximal separation power, which could be reached by adding a bias term, as shown in Theorem 5.2.5.

The theorem shows that the zero locus of a neural space of depth $d$ can be recursively computed as a combination of unions and intersections of the zero loci of neural spaces of depth $d-1$. At depth 1, the neural space reduces to a subspace of affine maps, and finding its zero locus corresponds to solving a system of linear equations. Although the actual execution of Formula 5.5 requires superpolynomial time, this recursive approach is particularly useful for deriving key properties of the separation relation, such as the role of activations, depth, and hidden features on the separation power, as detailed in the following sections.

### 5.2.3 The Role of Activations

The following result shows that the choice of the activation function—and its properties, such as injectivity or monotonicity—is irrelevant to separability, as long as the activation is non-polynomial.

**Theorem 5.2.5.** *Let $M_1, \dots, M_d$ be layer spaces with full bias and let $\sigma$ and $\tau$ be continuous activation functions, where $\sigma$ is non-polynomial. Then*

$$\rho\big(\mathcal{N}_\sigma(M_1, \dots, M_d)\big) \subseteq \rho\big(\mathcal{N}_\tau(M_1, \dots, M_d)\big).$$

*If $\tau$ is also non-polynomial, equality holds. Thus, non-polynomial activations not only yield equivalent separability but also achieve maximal separation power.*

*Proof.* Thanks to Proposition 5.2.1, we can work with zero loci $\mathcal{V}$ rather than with the separation relation $\rho$. We prove that non-polynomial activation functions have equivalent separation power by induction on $d$. If $d = 1$ then $\mathcal{V}(\mathcal{N}_\sigma(M_1)) = \mathcal{V}(M_1)$ which does not depend on $\sigma$. Now suppose that the zero locus $\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-1}))$ does not depend on $\sigma$ for each sequence of layer spaces $M_1, \dots, M_{d-1}$. Then, observing (5.5)

$$\mathcal{V}\big(\mathcal{N}_\sigma(M_1, \dots, M_d)\big) = \bigcap_{h,k} \bigcup_{\mathcal{P} \in \Psi_{h,k}} \bigcap_{\substack{P \in \mathcal{P} \\ i,j \in P}} \mathcal{V}\big(\mathcal{N}_\sigma(M_1, \dots, M_{d-2}, (M_{d-1})_{ij})\big),$$

we note that $\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_d))$ is independent of $\sigma$ as indices such as $h, k$ and $i, j$ are independent of $\sigma$, as well as $\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-2}, (M_{d-1})_{ij}))$ is by inductive hypothesis.

Finally, the first part of Theorem 5.2.5 follows directly from the proof of Theorem 5.2.2 and the last part of Theorem C.1.1. □

While non-polynomiality is sufficient for maximal separation, some polynomial activations may also achieve maximal separation power. Identifying these polynomials—or simply some of their properties, such as their degree—remains a complex mathematical problem. For more details, see [62].

### 5.2.4 The Role of Depth

Depth is a key hyperparameter influencing the separation power of neural spaces. Theorem 5.2.6 shows that, while adding layers of the same type can initially enhance separation power, this effect stabilizes after a finite number of layers.

**Theorem 5.2.6.** *Let $M_i$ be a layer space with complete bias in $\mathrm{Aff}_G(V_{i-1}, V_i)$ for each $i = 1, \dots, d$. Suppose that $V_{h-1} = V_h$ for some integer $1 \leq h \leq d$ and that $id_{V_h} \in M_h$. If $\sigma$ is a continuous non-polynomial activation function, then for $m \leq n$,*

$$\begin{aligned}\rho(\mathcal{N}_\sigma(M_1, \dots, M_{h-1}, \underbrace{M_h, \dots, M_h}_{n \text{ times}}, M_{h+1}, \dots, M_d)) \subseteq \\ \subseteq \rho(\mathcal{N}_\sigma(M_1, \dots, M_{h-1}, \underbrace{M_h, \dots, M_h}_{m \text{ times}}, M_{h+1}, \dots, M_d)),\end{aligned}$$

*but there exists a repetition threshold $R$ such that for all $n, m \geq R$ the inclusion becomes an equality.*

*Proof.* To prove the first part of the statement, by Lemma D.2.4, it suffices to show that

$$\mathcal{N}_\sigma(M_1,\dots,M_i,\dots,M_d) \subseteq \mathcal{N}_\sigma(M_1,\dots,\underbrace{M_i,\dots,M_i}_{n\text{-times}},\dots,M_d). \tag{5.13}$$

for each $n \geq 1$.

By recursion, it suffices to show the inclusion for $n = 1$ and $m = 2$. Moreover, Theorem 5.2.5 implies that it is sufficient to prove this inclusion for a single non-polynomial $\sigma$. Therefore, let $\sigma$ be the ReLU activation function, noting that in this case $\sigma \circ \sigma = \sigma$; equivalently

$$\tilde{\sigma} = \tilde{\sigma} \circ \tilde{\sigma} = \tilde{\sigma} \circ id_{\mathbb{R}^{X_i}} \circ \tilde{\sigma}.$$

Thus, each neural network

$$\begin{aligned}\phi_d \circ \tilde{\sigma} \circ \cdots \circ \tilde{\sigma} \circ \phi_1 \\ \in \mathcal{N}_\sigma(M_1,\dots,M_h,\dots,M_d).\end{aligned}$$

can be written as

$$\begin{aligned}\phi_d \circ \cdots \circ \tilde{\sigma} \circ id_{\mathbb{R}^{X_i}} \circ \tilde{\sigma} \circ \cdots \circ \tilde{\sigma} \circ \phi_1 \\ \in \mathcal{N}_\sigma(M_1,\dots,M_h,M_h,\dots,M_d).\end{aligned}$$

which is an element of, thereby proving the inclusion (5.13) by Lemma D.2.4.

The final step is to prove the stabilization property. This is achieved by recalling that, by Proposition 5.2.1 and Theorem 5.2.2,

$$\begin{aligned}&\mathcal{V}(\mathcal{N}_\sigma(\Psi(M_1),\Pi(M_2),\dots,\Pi(M_{d-1}),\mathrm{T}(M_d))) = \\ &\qquad \bigcap_{h,k} \bigcup_{\mathcal{P}\in\Psi_{h,k}} \bigcap_{\substack{P\in\mathcal{P}\\ i,j\in P}} \mathcal{V}(\mathcal{N}_\sigma(\Psi(M_1),\Pi(M_2),\dots,\Pi(M_{d-1}),\mathrm{T}(M_d))).\end{aligned} \tag{5.14}$$

Define

$$C_n := \mathcal{V}\big(\mathcal{N}_\sigma(\Psi(M_1),\Pi(M_2),\dots,\underbrace{\Pi(M_i),\dots,\Pi(M_i)}_{n\text{ times}},\dots,\Pi(M_{d-1}),\mathrm{T}(M_d))\big)$$

for each $n \in \mathbb{N}$. Recursively applying (5.14), both $C_n$ and $C_m$ can be represented as unions and intersections of elements in the finite set

$$\mathcal{C} = \{\mathcal{V}(\mathcal{N}_\sigma(\Psi(M_1),\Pi(M_2),\dots,\Pi(M_{i-1}),(\Pi(M_i)_{hk}))\}_{h,k\in X_i}.$$

We can reformulate the descending sequence (5.13) as follows

$$\cdots \subseteq C_n \subseteq C_{n-1} \subseteq \cdots \subseteq C_1$$

which stabilizes due to Lemma D.2.1. □

The repetition threshold may vary depending on the model and representation. For example, $k$-IGNs, being equivalent to $k$-WL, have a repetition threshold proportional to that of $k$-WL itself [77, 41]. In contrast, the following proposition shows an example of stabilization after just one repetition.

**Proposition 5.2.7.** *When the hidden representation spaces are regular representations, stabilization occurs after one layer repetition. Namely,*

$$\rho(\mathcal{N}_\sigma(V, \mathbb{R}^G, \dots, \mathbb{R}^G, \mathbb{R}^{G/H})) = \rho(\mathcal{N}_\sigma(V, \mathbb{R}^G, \mathbb{R}^{G/H})).$$

### 5.2.5 The Role of Intermediate Representations

In this section, we show that if a representation, $V$ can be decomposed as $V' \oplus V''$, then the separation power of neural spaces with hidden representation $V$ reduces to the combined separation power of two distinct neural spaces with hidden representations $V'$ and $V''$. In this section, we present Theorem 5.2.8, an additional application of Theorem 5.2.2, which demonstrates that the separation equivalence relation for neural spaces defined on $V$ is the intersection of those for neural spaces defined on $V'$ and $V''$. This implies that by decomposing each hidden representation V into a sum of *minimal* factors, the study of the separation power of general neural spaces can be reduced to analyzing those defined on minimal representations, as will be explored in Section 5.2.6.

For now, we focus on developing the notation necessary to state and prove Theorem 5.2.8. The structure of our network of interest is as follows:

$$\eta : V \xrightarrow{\phi_1} V_1 \xrightarrow{\tilde{\sigma}} \cdots \xrightarrow{\phi_i} V_i' \oplus V_i'' \xrightarrow{\tilde{\sigma}} V_i' \oplus V_i'' \xrightarrow{\phi_{i+1}} \cdots \xrightarrow{\tilde{\sigma}} V_d \xrightarrow{\phi_{d+1}} W$$

with $\eta \in \mathcal{N}_\sigma(M_1, \dots, M_d)$.

To formulate the separation equivalence relation for these networks in terms of the separation relations of simpler architectures with only $V'$ and $V''$ as intermediate representations, we define the projection and immersion maps as follows.

$$\pi' : V' \oplus V'' \to V', \quad \text{and} \quad \iota' : V' \to V' \oplus V''.$$

Similarly, we can define $\pi''$ and $\iota''$. Furthermore, for any $G$-representation $W$, we define

$$\pi'_* : \begin{array}{ccc} \mathrm{Aff}_G(W, V' \oplus V'') & \to & \mathrm{Aff}_G(W, V') \\ f & \mapsto & \pi' \circ f \end{array}$$

and

$$\iota'^* : \begin{array}{ccc} \mathrm{Aff}_G(V' \oplus V'', W) & \to & \mathrm{Aff}_G(V', W) \\ f & \mapsto & f \circ \iota' \end{array}$$

Similarly, we define $\pi''_*$ and $\iota''^*$. Let $M$ be a subspace of $\mathrm{Aff}_G(W, V' \oplus V'')$, its image $\pi'_*(M)$ is a subspace of $\mathrm{Aff}_G(W, V')$ and

$$M = \pi'_*(M) + \pi''_*(M).$$

Indeed, each $f \in M$ can be expressed as

$$f = \pi' f + \pi'' f = \pi'_*(f) + \pi''_*(f),$$

identifying $V'$ and $V''$ as subspaces of $V$. Similarly, for $M \subseteq \mathrm{Aff}_G(V' \oplus V'', W)$,

$$M = \iota'^*(M) + \iota''^*(M)$$

Hence, we can write

$$\mathcal{N}_\sigma(M_1, \ldots, M_d) = \\ \mathcal{N}_\sigma(M_1, \ldots, \pi'_*(M_i) + \pi''_*(M_i), \iota'^*(M_{i+1}) + \iota''^*(M_{i+1}), \ldots, M_d),$$

and the problem informally stated above reduces to determining the separation power of the entire family $\mathcal{N}_\sigma(M_1, \ldots, M_d)$ by understanding the separation power of the smaller families $\mathcal{N}_\sigma(M_1, \ldots, \pi' M_i, \iota'^*(M_{i+1}), \ldots, M_d)$ and $\mathcal{N}_\sigma(M_1, \ldots, \pi'' M_i, \iota''^*(M_{i+1}), \ldots, M_d)$. This is achieved by the following theorem.

**Theorem 5.2.8.** *Let $M_1, \ldots, M_d$ be layer space with complete bias. With the notation defined above, we have*

$$\rho\big(\mathcal{N}_\sigma(M_1, \ldots, M_d)\big) \\ = \rho\Big(\mathcal{N}_\sigma\big(M_1, \ldots, \pi'_*(M_i), \iota'^*(M_{i+1}), \ldots, M_d\big)\Big) \cap \\ \rho\Big(\mathcal{N}_\sigma\big(M_1, \ldots, \pi''_*(M_i), \iota''^*(M_{i+1}), \ldots, M_d\big)\Big).$$

*Remark* 5.2.9. Let $M$ and $N$ be complete layer spaces respectively in $\mathrm{Aff}_G(V, W)$ and $\mathrm{Aff}_G(W, U)$. Recalling Definition 4.3.8 of block spaces, we obtain

$$M^{1\times 2} = \mathrm{Aff}_G(V, W \oplus W) \quad \text{and} \quad N^{1\times 2} = \mathrm{Aff}_G(W \oplus W, U),$$

and then

$$\pi'_* M^{1\times 2} = \pi''_* M^{1\times 2} = M \quad \text{and} \quad \iota'^* N^{2\times 1} = \iota''^* N^{2\times 1} = N.$$

Thus, Theorem 5.2.8 implies

$$\begin{aligned} &\rho\big(\mathcal{N}_\sigma(M_1, \ldots, M_i^{1\times 2}, M_{i+1}^{2\times 1}, \ldots, M_d)\big) \\ &\quad = \rho\Big(\mathcal{N}_\sigma\big(M_1, \ldots, \pi'_*(M_i^{1\times 2}), \iota'^*(M_{i+1}^{2\times 1}), \ldots, M_d\big)\Big) \cap \\ &\qquad \rho\Big(\mathcal{N}_\sigma\big(M_1, \ldots, \pi''_*(M_i^{1\times 2}), \iota''^*(M_{i+1}^{2\times 1}), \ldots, M_d\big)\Big) \\ &\qquad\quad = \rho\Big(\mathcal{N}_\sigma\big(M_1, \ldots, M_i, M_{i+1}, \ldots, M_d\big)\Big) \cap \\ &\qquad\qquad \rho\Big(\mathcal{N}_\sigma\big(M_1, \ldots, M_i, M_{i+1}, \ldots, M_d\big)\Big) \\ &\qquad\qquad\qquad = \rho\Big(\mathcal{N}_\sigma\big(M_1, \ldots, M_d\big)\Big). \end{aligned} \tag{5.15}$$

As a result, applying the above recursively, for arbitrary $k_1, \ldots, k_{d-1}$ in $\mathbb{N}_{>0}$, we obtain

$$\rho\big(\mathcal{N}_\sigma(M_1^{1\times k_1}, M_2^{k_1\times k_2}, \ldots, M_{d-1}^{k_{d-2}\times k_{d-1}}, M_d^{k_{d-1}\times 1})\big) = \\ \rho\big(\mathcal{N}_\sigma(M_1, \ldots, M_d)\big).$$

Thus, the separability is independent of multiplicity and invariant features in intermediate representations.

Remark 5.2.9 implies the following result on the separation power of universality classes.

**Corollary 5.2.10.** *For layer spaces $M_1, \dots, M_d$ with complete bias, the separation relation of a neural space is the same of the associated universality class. Namely,*

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

### 5.2.6 The Role of Representation Type

Thanks to Theorem 5.2.8, we can focus on studying the separation power of neural spaces defined on *minimal* representations. These minimal representations are of the form $\mathbb{R}^X$, where the group $G$ acts transitively on $X$. That is, for any pair of points $x, y \in X$, there exists an element $g \in G$ such that $gx = y$. Basic group theory [39] shows that a set with a transitive action is in bijective correspondence with right cosets $G/H$ for some subgroup $H < G$, see Definition 2.1.5. Informally, the following theorem allows us to compare representations induced by transitive actions arising from certain subgroups. To simplify notation, for the remainder of this manuscript we will often abuse notation as follows. In the case where $M_i = \mathrm{Aff}_G(V_{i-1}, V_i)$ for each $i = 1, \dots, d$, we write

$$\mathcal{N}_\sigma(V_0, \dots, V_d) := \mathcal{N}_\sigma(M_1, \dots, M_d). \tag{5.16}$$

**Theorem 5.2.11.** *Let $K < H < G$ be finite groups. We have*

$$\rho(\mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/K}, \dots, W)) \subseteq \rho(\mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/H}, \dots, W)).$$

*Proof of Theorem 5.2.11.* Write $H/K = \{h_1 K, \dots, h_s K\}$, we have the following injection

$$\begin{aligned} \iota : \quad & \mathbb{R}^{G/H} \longrightarrow \mathbb{R}^{G/K} \\ & e_{gH} \mapsto \frac{1}{s} \sum_{i=1}^{s} e_{gh_i K} \end{aligned} \tag{5.17}$$

and projection

$$\begin{aligned} \pi : \quad & \mathbb{R}^{G/K} \longrightarrow \mathbb{R}^{G/H} \\ & e_{gK} \mapsto e_{gH}. \end{aligned} \tag{5.18}$$

Note that $\pi\iota = id_{\mathbb{R}^{G/H}}$, indeed,

$$\pi\iota(e_{gH}) = \frac{1}{s} \sum_{i=1}^{s} \pi(e_{gh_i K}) = \frac{1}{s} \sum_{i=1}^{s} e_{gh_i H} = e_{gH},$$

as $gh_i H = gH$ for each $i = 1, \dots, s$.

Consider the following diagram

$$\begin{array}{ccccccccc} \eta : V & \longrightarrow \cdots \xrightarrow{\phi} & \mathbb{R}^{G/H} & \xrightarrow{\sigma_H} & \mathbb{R}^{G/H} & \xrightarrow{\psi} & \cdots \longrightarrow & W \\ & & \pi \uparrow \downarrow \iota & & \pi \uparrow \downarrow \iota & & & \\ \eta' : V & \longrightarrow \cdots \xrightarrow{\phi'} & \mathbb{R}^{G/K} & \xrightarrow{\sigma'_H} & \mathbb{R}^{G/K} & \xrightarrow{\psi'} & \cdots \longrightarrow & W. \end{array}$$

From the network $\eta$ in $\mathcal{N}_\sigma(V, \ldots, \mathbb{R}^{G/H}, \ldots, W)$ composed by $\phi$, $\psi$, and $\sigma$ we want construct a new representation $\eta'$ defined as follows. Let $\phi' = \iota \circ \phi$, $\psi' = \psi \circ \pi$, and $\tilde{\sigma}' = \iota \circ \tilde{\sigma} \circ \pi$ and note that $\psi' \circ \tilde{\sigma}' \circ \phi' = \psi \circ \pi \circ \iota \circ \tilde{\sigma} \circ \pi \circ \iota \circ \phi = \psi \circ \tilde{\sigma} \circ \phi$. Hence, substituting $\psi \circ \tilde{\sigma} \circ \phi$ with $\psi' \circ \tilde{\sigma}' \circ \phi'$ inside the definition of $\eta$ do not change the function, and embeds it into a parameter space with intermediate representation $\mathbb{R}^{G/K}$ instead of $\mathbb{R}^{G/H}$. But to prove that $\eta$ is a neural network, we need to prove that $\tilde{\sigma}'$ is a point-wise activation function for some real-valued function $\sigma'$.

If $\tilde{\sigma}$ is a point-wise activation associated to $\sigma : \mathbb{R} \to \mathbb{R}$ defined on $\mathbb{R}^{G/H}$ we have that

$$\tilde{\sigma}(\sum_{gH \in G/H} a_{gH} e_{gH}) = \sum_{gH \in G/H} \sigma(a_{gH}) e_{gH}.$$

On the other hand, we have

$$\tilde{\sigma}'(\sum_{gK \in G/K} a_{gK} e_{gK}) = \iota \circ \tilde{\sigma} \circ \pi(\sum_{gK \in G/K} a_{gK} e_{gK}) =$$

$$\iota \circ \tilde{\sigma}(\sum_{\substack{gH \in G/H \\ ghK \in gH/K}} a_{ghK} e_{gH}) = \iota \sum_{gH \in G/H} \sigma(\sum_{ghK \in gH/K} a_{ghK}) e_{gH} =$$

$$\frac{1}{s} \sum_{gH \in G/H} \sigma(\sum_{ghK \in gH/K} a_{ghK}) \sum_{hK \in H/K} e_{ghK} =$$

$$\frac{1}{s} \sum_{gK \in G/K} \sigma(\sum_{hK \in H/K} a_{ghK}) e_{gK}.$$

Note that the map

$$\alpha : \sum_{gK \in G/K} a_{gK} e_{gK} \mapsto \sum_{hK \in H/K} a_{ghK} e_{gK}$$

is linear and $G$-equivariant. In particular, note that $\tilde{\sigma}' = \frac{\tilde{\sigma}_K \circ \alpha}{s}$, where we denote the standard point-wise activation induced by $\sigma$ on $\mathbb{R}^{G/K}$ as $\tilde{\sigma}_K$, to distinguish it from $\tilde{\sigma}$, the point-wise activation induced by $\sigma$ but defined on $\mathbb{R}^{G/H}$. Hence, substituting

$$\psi \circ \tilde{\sigma} \circ \phi$$

with

$$\psi' \circ \tilde{\sigma}' \circ \phi' = \psi' \circ \frac{\tilde{\sigma}_K \circ \alpha}{s} \circ \phi',$$

we obtain an immersion of $\eta$ in $\mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/K}, \dots, W)$. Hence

$$\mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/H}, \dots, W) \subseteq \mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/K}, \dots, W)$$

and

$$\rho(\mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/K}, \dots, W)) \subseteq \rho(\mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/H}, \dots, W)).$$

□

Theorem 5.2.11 implies that neural spaces with minimal representations in one layer, namely $\{\mathcal{N}_\sigma(V, \dots, \mathbb{R}^{G/H}, \dots, W)\}_{H<G}$, form a separation power hierarchy corresponding to the hierarchy of subgroups of $G$. In particular, if $H = G$, the corresponding representation $\mathbb{R}^{G/G}$ has minimal separation power. Furthermore, notice that $\mathbb{R}^{G/G} \cong \mathbb{R}$ is the trivial representation, and this means that invariant layers have the lowest separation power. On the other hand, if $H = \{e\}$ is the group containing only the identity element, the corresponding representation $\mathbb{R}^{G/\{e\}}$ has maximal separation power, since $\{e\}$ is contained in every subgroup of $G$. Since $\mathbb{R}^{G/\{e\}} \cong \mathbb{R}^G$ is the regular representation, this implies that the regular representation achieves the maximum separation power. In general, if $K < H$, then $\dim \mathbb{R}^{G/H} < \dim \mathbb{R}^{G/K}$. Hence, improving separability requires working in a larger space, which, aside from ad-hoc optimizations, leads to additional computational cost. In particular, by applying Theorem 5.2.2, we can prove the following proposition.

**Proposition 5.2.12.** *The neural space $\mathcal{N}_\sigma(V, \mathbb{R}^G, \mathbb{R}^{G/H})$ of equivariant shallow networks with regular hidden representations identifies inputs if and only if they belong to the same $H$-orbit, i.e., $(\beta, \beta') \in \rho(\mathcal{N}_\sigma(V, \mathbb{R}^G, \mathbb{R}^{G/H}))$ if and only if there exists some $h \in H$ such that $h\beta = \beta'$.*

This is consistent with the results in [99], which demonstrate that shallow networks with hidden representation blocks isomorphic to $\mathbb{R}^G$ are separation-constrained universal with respect to maximal separation power, as stated in Theorem 16 of [54].

## 5.3 Implications on Specific Models

### 5.3.1 Invariant Graph Networks

Theorem 1 in [77] and Theorem 2 in [41] together imply the following result for the theory of IGNs. Employing again the convention introduced in (5.16).

**Proposition 5.3.1.** *There exist $d > 0$ and a **large** $F > 0$ such that for hidden feature dimensions $f_1, \dots, f_d > F$, the neural space*

$$\mathcal{N}_\sigma((\mathbb{R}^n)^{\otimes 2} \otimes \mathbb{R}^{f_0}, (\mathbb{R}^n)^{\otimes k} \otimes \mathbb{R}^{f_1}, \dots, (\mathbb{R}^n)^{\otimes k} \otimes \mathbb{R}^{f_d}, \mathbb{R})$$

*matches the separation power of $k$-WL.*

However, Remark 5.2.9 shows that the dimension of hidden invariant features does not affect separation power, strengthening Proposition 5.3.1 in the following corollary.

**Corollary 5.3.2.** *There exist $d > 0$ such that for **any** hidden feature dimensions $f_1, \dots, f_d > 0$, the neural space*

$$\mathcal{N}_\sigma((\mathbb{R}^n)^{\otimes 2} \otimes \mathbb{R}^{f_0}, (\mathbb{R}^n)^{\otimes k} \otimes \mathbb{R}^{f_1}, \dots, (\mathbb{R}^n)^{\otimes k} \otimes \mathbb{R}^{f_d}, \mathbb{R})$$

*matches the separation power of $k$-WL.*

### 5.3.2 Convolutional Neural Networks

The separation power of circular CNNs is influenced by the width of the filter's support.

**Proposition 5.3.3.** *Let $C^k$ be the layer space for circular convolutions with filter size $k$, as defined in Example 4.2.7. Consider the neural space*

$$k\text{-}CNN = \mathcal{N}_\sigma(C^k, \mathrm{Aff}_{\mathbb{Z}_n}(\mathbb{R}^n, \mathbb{R})).$$

*This is the space associated with shallow convolutional networks, where the first layer consists of one filter of size $k$ followed by an output invariant layer. For $n > 2$, we have:*

$$\rho(n\text{-}CNN) \subsetneq \rho(1\text{-}CNN), \quad \text{and} \quad \rho(n\text{-}CNN) \subseteq \cdots \subseteq \rho(2\text{-}CNN) \subseteq \rho(1\text{-}CNN).$$

# Chapter 6

# Shallow Equivariant Networks

In Chapter 5 we presented tools to understand the separation power of equivariant neural networks. However, 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 6.1). 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 6.1.1).

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 6.2). 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 [98] produces a broad family of shallow models that are universal within the separation-constrained function class (Section 6.3). 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 contributions of this chapter as follows:

- We characterize the universality classes of shallow invariant networks (Theorem 6.1.2).
- We establish general *sufficient conditions* under which universality fails, even within function classes exhibiting maximal separation (Theorem 6.1.3 and Theorem 6.1.4).
- 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 6.2.2).
- We generalize the results by [98] to a broader family of models (Theorem 6.3.2).

## 6.1 Characterization of Universality Classes

In this section, we characterize the universality classes of invariant shallow neural networks and compare them in particular cases. To this end, we begin by introducing the notion of a *basis map*.

**Definition 6.1.1** (Basis maps)**.** 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 $\ell$, which we identify with $[\ell]$. Let $\phi^1, \dots, \phi^m$ be a basis for the linear part of $M$, and for each $i \in Y$, define the linear maps

$$\phi_i : \begin{array}{c} \mathbb{R}^X \to \mathbb{R}^m \\ x \mapsto (\phi_i^1(x), \dots, \phi_i^m(x)). \end{array} \tag{6.1}$$

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

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

**Theorem 6.1.2.** *Let $M$ and $N$ be, respectively, layer spaces with complete bias in $\mathrm{Aff}_G(V, W)$ and $\mathrm{Aff}_G(W, \mathbb{R})$. Let $f$ be an invariant function, then $f$ belongs to $\mathcal{U}_\sigma(M, N)$ 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 $\phi_i^1, \ldots, \phi_i^m$ of each basis map $\phi_1, \ldots, \phi_\ell$, see* (6.1).

Here, we assume $d = \dim V$, and let $P(\partial_1, \ldots, \partial_d)$ denote the constant-coefficient linear differential operator associated with the polynomial $P$. The derivatives $\partial_i$ on $V$ are interpreted in the distributional sense; see [47] for details.

Although Theorem 6.1.2 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 $\phi_i^1, \ldots, \phi_i^m$. 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.

### 6.1.1 Sufficient Conditions for Universality Failure

In this section, we present two sufficient conditions for the failure of separation-constrained universality. We begin with Theorem 6.1.3, which provides a general—but more difficult to verify—criterion, followed by Theorem 6.1.4, 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 $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$, the directional derivative is defined as the differential operator $D_c = c_1 \cdot \partial_1 + \cdots + c_n \cdot \partial_n$.

**Theorem 6.1.3.** *A continuous function $f$ does not belong to the class $\mathcal{U}_\sigma(M, N)$ if*

$$D_{c_1} \cdots D_{c_\ell} f \neq 0 \tag{6.2}$$

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

In the case of equivariant networks where each affine layer is allowed to be an arbitrary equivariant affine map, Theorem 6.1.3 can be strengthened as follows.

**Theorem 6.1.4.** *Let $M = \mathrm{Aff}_G(V, W)$ and $N = \mathrm{Aff}_G(W, \mathbb{R})$, where $V$ and $W$ are permutation representations. Let $\phi_1, \ldots, \phi_\ell$ denote the basis maps associated with $M$, see* (6.1). *Then, the universal class $\mathcal{U}_\sigma(M, N)$ fails to be separation-constrained universal if, for some choice of:*

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

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

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

We now apply the tools developed here to investigate the structure of universality classes and illustrate their heterogeneity. In Section 6.2, we apply Theorems 6.1.3 and 6.1.4 to exhibit concrete examples of failure. In contrast, Section 6.3 presents Theorem 6.3.2, a generalization of Theorem 4.4.2, which provides sufficient conditions for achieving separation-constrained universality—highlighting the diversity of behaviors even within fixed symmetry classes.

## 6.2 Examples of Failure

In the following proposition, we identify three universality classes associated with different architectures but with the same separation power.

**Proposition 6.2.1.** *Let $C$ be defined as in Example 4.2.7, representing convolutional filters of width 1, let $I$ be the layer space defined as in Example 4.2.5, representing invariant circular layers, and let $P$ be the layer space of equivariant affine maps defined in Example 4.2.6. Let $S_n$ act on $\mathbb{R}^n \cong \mathbb{R}^{[n]}$ via the standard permutation action, and on $\mathbb{R}^{S_n}$ via the regular representation. Then the following universality classes have the same separation power:*

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

*In the last term we use the notation introduced in* (5.16).

However, the universality classes appearing in Proposition 6.2.1 have the same separation power but are different, as shown in Proposition 6.2.2.

**Proposition 6.2.2.** *As established in Proposition 6.2.1, the following spaces achieve the same separation power, yet differ in their approximation capabilities when $n > 2$:*

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

We will prove the two strict inclusions of Proposition 6.2.2 in the following three paragraphs.

**Failure for CNN with filter width 1:** We now apply Theorem 6.1.3 to show that CNNs with filter width 1 cannot approximate the function $(x_1+\cdots+x_n)^n$ for $n>1$, namely $(x_1+\cdots+x_n)^n \notin \mathcal{U}_\sigma(C,I)$. Indeed, for any $\alpha = 1,\dots,n$, we have $e_{\alpha+1} \in \ker(\pi_\alpha^\top) = \mathrm{Span}\{e_1,\dots,\hat{e}_\alpha,\dots,e_n\}$, 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 (6.2) in Theorem 6.1.3.

**Success for PointNet:** We now show that shallow PointNets approximate the polynomial function $(x_1+\cdots+x_n)^n$. By Proposition D.3.2 in Appendix D.3, $f(x_1,x_1+\cdots+x_n)+\cdots+f(x_n,x_1+\cdots+x_n)$ 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 6.2.2, namely $\mathcal{U}_\sigma(C,I) \subsetneq \mathcal{U}_\sigma(\mathbb{R}^n,\mathbb{R}^n,\mathbb{R})$.

**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 4.4.2.
We start considering ($n > 3$). We again employ Theorem 6.1.3 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:

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

Explicit computation shows that $D_{c_n} \cdots D_{c_1}(x_1 \cdots x_n) = 2$, verifying (6.2). The previous technique does not apply in the case $n = 3$, for which we must instead resort to Theorem 6.1.4. 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 [106], 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.

## 6.3 Examples of Success

We now present Theorem 6.3.2, a generalization of Theorem 4.4.2, which shows that a specific class of hidden representations can achieve separation-constrained universality. These representations arise from cosets of certain subgroups $H$ of $G$, which we define as follows:

**Definition 6.3.1** (Normal subgroup)**.** A subgroup $H$ is *normal* if $ghg^{-1} \in H$ for each $h \in H, g \in G$.

We state the next theorem, again employing the notation introduced in (5.16).

**Theorem 6.3.2.** *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* 6.3.3. 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 6.2.2.

# Chapter 7

# Deep Equivariant Networks

Chapter 6 suggests a more nuanced landscape for the interaction between separation and universality. Indeed, we presented 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, [124], [95], and [106] 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 chapter, 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. The first result presented here is a *separation-constrained universality theorem* for invariant networks, showing that models with fully connected readouts can approximate every continuous function that is consistent with their separation relation (Section 7.1). We then turn to the equivariant setting, where a simple example shows that standard separability is too coarse to characterize universality. With this notion in place, we prove two *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 7.2). 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 7.1.1), showing that the addition of a fully connected readout guarantees approximation within the separation-constrained class.
- We prove two separation-constrained universality theorems for the *fully equivariant* case, showing that equivariant networks achieve universality either once the separation relations stabilize with depth (Theorem 7.2.1), or when equipped with specific readout layers (Theorem 7.2.3).

## 7.1 Universality of Invariant Neural Networks

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

**Theorem 7.1.1.** *Let $M_1, \dots, M_d$ be layer spaces with full bias as defined in Definition 4.2.2 and let $I$ be a layer space of invariant affine functions. Set*

$$\rho = \rho\big(\mathcal{U}_\sigma(M_1, \dots, M_d, I)\big).$$

*Then*

$$\mathcal{U}_\sigma(M_1, \dots, M_d, I, L) = \mathcal{C}_\rho(V). \tag{7.1}$$

*Proof.* Note that by Corollary 5.2.10, the relation $\rho$ is preserved under the extension from the neural space $\mathcal{N}_\sigma(M_1, \dots, M_d, I, L)$ to the universality class $\mathcal{U}_\sigma(M_1, \dots, M_d, I, L)$ and from $\mathcal{N}_\sigma(M_1, \dots, M_d, I)$ to $\mathcal{U}_\sigma(M_1, \dots, M_d, I)$, therefore

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

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

$$\mathcal{A}_h := \Big\{\eta \circ F_h \mid \eta \in \mathcal{C}(\mathbb{R}^h)\Big\}, \quad \mathcal{A}'_h := \big\{\eta \circ F_h \mid \eta \in \bigcup_{k \in \mathbb{N}} \mathcal{N}_\sigma(L^{h\times k}, L^{k\times 1})\big\}. \tag{7.2}$$

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

$$\mathcal{C}_\rho(V) \overset{\text{Lemma } D.4.4}{=} \overline{\bigcup_{h\in\mathbb{N}} \mathcal{A}_h} = \overline{\bigcup_{h\in\mathbb{N}} \mathcal{A}'_h}. \tag{7.3}$$

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

$$\begin{aligned}
\mathcal{U}_\sigma(M_1,\dots,M_d,I,L) &= \overline{\bigcup_{\vec{k}\in\mathbb{N}^{d+1}} \mathcal{N}_\sigma(M_1^{1\times k_1},\dots,M_d^{k_{d-1}\times k_d}, I^{k_d\times k}, L^{k\times 1})} \\
&= \overline{\bigcup_{\tilde{k}\in\mathbb{N}^{d+2}} \mathcal{N}_\sigma(M_1^{1\times k_1},\dots,M_d^{k_{d-1}\times k_d}, I^{k_d\times h}) \,\hat{\circ}\, \mathcal{N}_\sigma(L^{h\times k}, L^{k\times 1})} \\
&= \overline{\bigcup_{h\in\mathbb{N}} \left\{ \eta\circ f \mid f\in\mathcal{N}_h,\ \eta\in \bigcup_{k\in\mathbb{N}} \mathcal{N}_\sigma(L^{h\times k}, L^{k\times 1}) \right\}} \\
&\overset{\text{Equation 7.2}}{\supseteq} \overline{\bigcup_{h\in\mathbb{N}} \mathcal{A}'_h} \overset{\text{Equation 7.3}}{=} \mathcal{C}_\rho(V).
\end{aligned}$$

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 Proposition 4.3.7. The last equality holds because of Equation 7.3 and Corollary D.4.3, 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 $\mathcal{A}'_h$. □

## 7.2 Universality of Equivariant Neural Networks

In this section, we extend the previous results to the equivariant setting. We now state universality results under the more general notion of separability for arbitrary codomains, and we recall Definition 5.1.4.

**Theorem 7.2.1.** *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 with full bias in $\mathrm{Aff}_G(V_{i-1},V_i)$ for $i=1,\dots,f$, and let $M$ be a layer space with full bias in $\mathrm{Aff}_G(\mathbb{R}^X,\mathbb{R}^X)$ containing the identity map. Let $d$ be such that*

$$\rho := \rho\big(\mathcal{U}_\sigma(M_1,\dots,M_f,\underbrace{M,\dots,M}_{d \text{ times}})\big) = \rho\big(\mathcal{U}_\sigma(M_1,\dots,M_f,\underbrace{M,\dots,M}_{d+1 \text{ times}})\big). \tag{7.4}$$

*Then,*

$$\mathcal{U}_\sigma(M_1,\dots,M_f,\underbrace{M,\dots,M}_{d+1 \text{ times}}) = \mathcal{C}_\rho(V_0,\mathbb{R}^X).$$

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

*Proof.* For brevity, we will prove all subsequent results in terms of $\mathcal{U}_\sigma(\underbrace{M, ..., M}_{d \text{ times}})$ or similar forms. The same results, however, extend verbatim to the more general setting $\mathcal{U}_\sigma(M_1, \ldots, M_d, \underbrace{M, ..., M}_{d \text{ times}})$. Note that

$$\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d \text{ times}}, C) \subseteq \mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d+1 \text{ times}}).$$

Therefore,

$$\rho\big(\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d+1 \text{ times}})\big) \subseteq \rho\big(\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d \text{ times}}, C)\big) \subseteq \rho\big(\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d \text{ times}})\big).$$

By hypothesis,

$$\rho := \rho\big(\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d+1 \text{ times}})\big) = \rho\big(\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d \text{ times}})\big),$$

and hence

$$\rho = \rho\big(\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d \text{ times}}, C)\big)$$

as well. Then, by Theorem 7.2.3,

$$\mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d \text{ times}}, C) = \mathcal{C}_\rho(V).$$

Since functions realized by neural networks are continuous, it follows that

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

Finally, we observe that

$$\mathcal{C}_\rho(V) = \mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d \text{ times}}, C) \subseteq \mathcal{U}_\sigma(\underbrace{M, \ldots, M}_{d+1 \text{ times}}) \subseteq \mathcal{C}_\rho(V),$$

which yields the claim. □

Since by Theorem 5.2.6, separation is known to stabilize after a certain depth, we obtain the following corollary.

**Corollary 7.2.2.** *Assume the notation of Theorem 7.2.1. There exists a natural number $D$ for which $\mathcal{U}_\sigma(M_1, \ldots, M_f, \underbrace{M, ..., M}_{d \text{ times}})$ is separation-constrained universal for each $d \geq D$.*

In a different direction, we can show that 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 7.2.1 and Corollary 7.2.2. This is formalized as follows.

**Theorem 7.2.3.** *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 with full bias 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. Then $\mathcal{U}_\sigma(M_1, \dots, M_f, C) = \mathcal{C}_\rho(V)$, where $\rho := \rho\big(\mathcal{U}_\sigma(M_1, \dots, M_f, C)\big)$.*

*Proof.* Recall that

$$C \subseteq \mathrm{Aff}_G(\mathbb{R}^X, \mathbb{R}^X).$$

Thanks to Proposition D.4.5, to prove Theorem 7.2.3 it suffices to show that, for each $x \in X$,

$$\pi_{x^*}\, \mathcal{U}_\sigma(M_1, \dots, M_f, C) = \mathcal{C}_{\rho_x}(V),$$

where

$$\rho_x := \rho\big(\pi_{x^*}\, \mathcal{U}_\sigma(M_1, \dots, M_f, C)\big).$$

Fix $x \in X$, and define $P_x := \pi_{x^*} C$. Then

$$\begin{aligned} \pi_{x^*}\, \mathcal{U}_\sigma(M_1, \dots, M_f, C) &= \pi_{x^*}\, \mathcal{U}_\sigma(M_1, \dots, M_f, \pi_{x^*} C) \\ &= \mathcal{U}_\sigma(M_1, \dots, M_f, P_x) \\ &= \mathcal{U}_\sigma(M_1, \dots, M_{f-1}, \pi_{x^*} M_f, L). \end{aligned}$$

For the second equality, note that

$$\begin{aligned} P_x &= \pi_x^* C \\ &= \{\, v \mapsto \pi_x(\lambda v + \mu \mathbb{1}) \mid \lambda, \mu \in \mathbb{R} \,\} \\ &= \{\, v \mapsto \lambda \pi_x(v) + \mu \mid \lambda, \mu \in \mathbb{R} \,\}. \end{aligned}$$

For the third equality, recall that the pointwise activation $\tilde{\sigma}$ is defined by

$$\tilde{\sigma}\Big(\sum_{x \in X} v_x e_x\Big) := \sum_{x \in X} \sigma(v_x)\, e_x,$$

and observe that, for each $x \in X$, we have the commutation relation

$$\sigma \circ \pi_x = \pi_x \circ \tilde{\sigma}.$$

Hence, at the final activation we have

$$\{\, v \mapsto \lambda \pi_x(\tilde{\sigma}(v)) + \mu \mid \lambda, \mu \in \mathbb{R} \,\} = \{\, v \mapsto \lambda \sigma(\pi_x(v)) + \mu \mid \lambda, \mu \in \mathbb{R} \,\},$$

and note that, in this way, the final layer becomes the space of arbitrary affine maps of the real line, namely $L$. Finally, note that $M_f \subseteq \mathrm{Aff}_G(V, \mathbb{R}^X)$ for some permutation representation $V$. Thus, by Remark D.4.6, we have

$$\pi_{x^*} M_f \subseteq \mathrm{Aff}_{G_x}(V, \mathbb{R}),$$

and therefore $\pi_{x^*} M_f$ is a space of $G_x$-invariant affine functions. Thanks to Theorem 7.1.1, we obtain

$$\pi_{x^*}\, \mathcal{U}_\sigma(M_1, \dots, M_f, C) = \mathcal{U}_\sigma(M_1, \dots, M_{f-1}, \pi_{x^*} M_f, L) = \mathcal{C}_{\rho_x}(V),$$

where

$$\rho_x = \rho\big(\pi_{x^*}\, \mathcal{U}_\sigma(M_1, \dots, M_f, C)\big) = \rho\big(\mathcal{U}_\sigma(M_1, \dots, M_{f-1}, \pi_{x^*} M_f, L)\big).$$

This concludes the proof. □

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 7.2.3, which is formulated in the equivariant setting, specializes to Theorem 7.1.1—the corresponding result in the invariant setting.

At first sight, it may be tempting to compare Theorem 7.2.1 and Theorem 7.2.3 and conclude that Theorem 7.2.3 is a stronger statement. However, it is important to note that adding the $C$ layer space at the end does not change the separation power of the model class, whereas adding a certain number of $M$ layers may increase it. Theorem 7.2.1 explicitly accounts for this effect.

Theorem 7.2.1 and Corollary 7.2.2 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 7.2.3, on the other hand, is instrumental in recovering known results such as [106]. It also shows that universality stabilization in Theorem 7.2.1 and Corollary 7.2.2 can occur at the same depth as separation stabilization, revealing that the threshold in Theorem 7.2.1 is not always optimal.

*Remark* 7.2.4. Thanks to Theorem 7.2.3, we can easily recover the universality result of [106]. 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 D.4.10 (Appendix D.4.3).

Note that this shows that the depth threshold required for separation stability in Theorem 7.2.1 provides a *sufficient*, but not *necessary*, condition for universality. Indeed, $\rho\big(\mathcal{U}_\sigma(V, \mathbb{R}^G, \mathbb{R}^G, W)\big) = \rho\big(\mathcal{U}_\sigma(V, \mathbb{R}^G, W)\big)$, so separation has not stabilized, yet separation-constrained universality is already achieved. However, determining in general when separation stabilization takes place is a difficult problem. Corollary 7.2.2 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* 7.2.5. Theorem 7.2.3 marks a significant difference between the equivariant and the invariant cases. Indeed, Proposition 6.2.1 shows that

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

although 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 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 $Y$.

# Chapter 8

# Conclusions

We developed a mathematical account of how architectural choices affect approximation in equivariant neural networks. We focused on permutation representations and continuous point-wise activations, which provide a broad yet tractable framework capturing many widely used models. First, we analyzed the interaction between point-wise nonlinearities and equivariance, characterizing when such activations induce non-trivial equivariant nonlinear maps, and identifying when the resulting hypothesis spaces are non-degenerate. Second, we studied the separation power for families of equivariant architectures, and we characterized how hyperparameters and architectural choices affect the ability to distinguish non-equivalent inputs. Third, we studied universality constrained by separation, and we showed that separation alone does not fully determine the set of functions we can approximate. Finally, we establish separation-constrained universality results for broad architecture families, and we identify depth and suitable readouts as decisive mechanisms enabling universality within the separation-constrained regime.

While the framework is broad, the presented work has two main limitations.

- Our analysis is restricted to permutation representations and point-wise nonlinearities. This excludes large and practically important families of equivariant models that rely on non-pointwise nonlinearities, such as tensor-product and gated nonlinearities, norm nonlinearities, or other equivariant nonlinear constructions [114, 63, 119, 118].
- The thesis primarily established universality results, which are asymptotic in nature. We do not provide approximation rates quantifying how approximation error decreases with width, depth, or other architectural choices. As a consequence, the results do not directly translate into approximation error bounds for hypothesis spaces used in practice which have fixed width and depth.
- As anticipated in Section 1.1, universality in the uniform norm is stronger than what is required to achieve arbitrarily small approximation error.

Therefore, to better match supervised learning practice, the present framework should be extended to the $L^2$ norm, or to approximation error measured by risk functionals induced by losses that are not norms, such as the cross-entropy loss. A particularly interesting problem is to understand the role of separation in these settings.

This research line also suggests several directions for advancing beyond approximation theory and for developing frameworks more tightly descriptive of practices in equivariant deep learning. In particular, in Section 1.1, we introduced the following excess risk decomposition.

$$\begin{aligned} \mathcal{R}(\tilde{f}) - \mathcal{R}(f^\star) &= && \textbf{(Excess risk)} \\ &= \mathcal{R}(\bar{f}) - \mathcal{R}(f^\star) && \textbf{(Approximation error)} \\ &+ \mathcal{R}(\tilde{f}) - \mathcal{R}(\bar{f}). && \textbf{(Learning error)} \end{aligned}$$

The learning error can be further decomposed as follows.

$$\begin{aligned} \mathcal{R}(\tilde{f}) - \mathcal{R}(\bar{f}) &= && \textbf{(Learning error)} \\ &= \mathcal{R}(\hat{f}) - \mathcal{R}(\bar{f}) && \textbf{(Estimation error)} \\ &+ \mathcal{R}(\tilde{f}) - \mathcal{R}(\hat{f}) && \textbf{(Optimization error)} \end{aligned}$$

where we define

$$\mathrm{ERM}_S(\mathcal{H}) := \operatorname*{arg\,min}_{f \in \mathcal{H}} \hat{\mathcal{R}}_S(f) \quad \text{and} \quad \hat{f} \in \operatorname*{arg\,min}_{f \in \mathrm{ERM}_S(\mathcal{H})} \mathcal{R}(f).^1$$

With this definition, the estimation error corresponds to selecting the best predictor among the empirical risk minimizers in $\mathcal{H}$. In the noiseless setting, or as the sample size tends to infinity, $\hat{f}$ may coincide with the best-in-class predictor $\bar{f}$. Finally, the optimization error captures the mismatch between the best predictor among the empirical risk minimizers and the model actually selected by the training algorithm $\mathcal{A}$, such as gradient descent or its stochastic and adaptive counterparts. We recap below some open problems and future directions to better understand the role that equivariance may play in estimation and optimization error.

- **Estimation Error:** A natural next step is to develop a general theory to shed light on the generalization properties of the models studied in the presented framework and why those models, if matched with properly biased data, generalize better. In particular, current efforts in equivariant machine learning literature only involve sample complexity measures

[1] Note that this definition is not the standard one, in which $\hat{f}$ is simply chosen as an element of $\mathrm{ERM}_S(\mathcal{H})$. However, in the interpolation regime, $\mathcal{R}(\hat{f})$ may depend on the particular choice of $\hat{f}$, so the estimation error $\mathcal{R}(\hat{f}) - \mathcal{R}(\bar{f})$ is not well-defined. To address this issue in a simple way, we adopt this naive definition, selecting a population-risk minimizer among empirical risk minimizers. Alternative approaches exist in the literature, often based on canonical selectors on $\mathcal{H}$, or by defining the estimation error to subsume the entire learning error, see [52, 103, 45].

such as Rademacher complexity and VC dimension [8, 32], or worst-case generalization gaps [102, 129] which are notoriously not adequate to describe the generalization phenomena of models in the interpolation regime [126, 10].

- **Optimization Error:** The previous observations raise questions about the optimization dynamics of neural networks and the presence of implicit bias due to optimization. Current accounts of this problem form a landscape that remains incomplete and heterogeneous, even in the simple case of fully connected neural networks. The optimization dynamics are only understood in restricted cases such as deep linear networks [103, 20], diagonal linear networks [88], the large-width regime [71], or the kernel regime [51], which remain far from the models employed in practice [108, 107, 109]. Of particular interest, the presence of symmetries in weight space seems to affect the learning dynamics and, in turn, training [128, 127], suggesting that non-identifiability can be a benign feature of training [76]. Other approaches to the study of optimization dynamics are model-agnostic and focus solely on the dynamics [11, 3]. In addition, another part of the literature highlights the importance of the loss landscape, showing that implicit regularization occurs even when the optimization algorithm is not gradient-based [23, 1, 87]. Regarding equivariant architectures, equivariant versions of neural tangent kernels have been proposed [81], but only for networks with regular hidden representations, and hence without addressing the separation constraint. Interesting differences between standard neural networks and some specific equivariant ones in the loss landscape have been shown by [123]. In the deep linear case, [48] showed that the implicit bias of gradient descent differs between standard convolutional networks and fully connected networks.

This is only a small portion of what remains poorly understood regarding the role of architectural choices in neural networks and their effect on performance. The work presented in this manuscript is barely sufficient to scratch the surface of this problem.

# Appendix A

# Basic Notions on Commutative Algebra

For a general introduction to commutative algebra, we refer to **(author?)** [4]. Here, we recall the notation necessary to prove Theorem 6.1.2, 6.1.3 and 6.1.4.

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

**Definition A.0.1** (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 A.0.2** (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 A.0.3** (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 A.0.4.** *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* A.0.5. 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 $X_1^\perp, \ldots, X_\ell^\perp$.

# Appendix B

# Basic Notions on Ridge Functions

In this section, we present results on the theory of superpositions of generalized ridge functions. A detailed exposition can be found in **(author?)** [91].

**Definition B.0.1** (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 $\Omega$ 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 $a_i$s 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 [91], p. 65).

**Theorem B.0.2.** *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 B.0.3.** Given $\Omega \subseteq \mathbb{R}^n$, define the ideal of polynomials vanishing on $\Omega$ 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 B.0.4** (Theorem 6.9 of [91])**.** *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 B.0.5.** *Let $\Omega$ and $\Omega'$ 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 B.0.6, then $\overline{C(\Omega)} \subseteq \ker q(D)$ while $p' \notin \ker q(D)$, concluding the proof. □

**Lemma B.0.6.** *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 $\phi$ 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 $Q(\partial_1, \ldots, \partial_d)$ be the adjoint operator of $P(\partial_1, \ldots, \partial_d)$. This operator is still a linear differential operator when defined on smooth functions with compact support. In particular, $Q(\partial_1, \ldots, \partial_d)\phi$ is still a smooth function with support on $K$. Moreover,

$$\langle p_n, Q(\partial_1, \ldots, \partial_d)\phi\rangle = -\langle P(\partial_1, \ldots, \partial_d)p_n, \phi\rangle = 0 \tag{B.1}$$

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

$$\langle p_n, Q(\partial_1, \ldots, \partial_d)\phi\rangle \to \langle p, Q(\partial_1, \ldots, \partial_d)\phi\rangle, \tag{B.2}$$

for $n \to \infty$. Thanks to Eq. B.1 e B.2 we get:

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

Finally,

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

Since $\phi$ is an arbitrary smooth function with compact support, we get

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

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

# Appendix C

# Basic Notions on Functional Equations

In this section, we introduce key results from the theory of functional equations that are necessary to prove Theorem 3.1.3. A functional equation, by definition, is an identity involving unknown functions as variables, and common examples include differential and integral equations [58]. Here, we are particularly interested in the class of linear functional equations, which we explore in greater detail in the following section.

## C.1 Linear Functional Equations

Linear functions equations are functional equations which, for given $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}^d$, are defined by

$$\sum_{i=1}^{n} a_i \sigma(b_i x) = 0 \qquad (\forall x \in \mathbb{R}^d).$$

In particular, Theorem C.1.1 is a fundamental tool in the proof of Theorem 3.1.3, since it characterizes the set of parameters $b_1, \ldots, b_n$ for which the specific case of linear functional equation in (C.1) is always satisfied for a non-polynomial $\sigma$ and arbitrary $a_1, \ldots, a_n \in \mathbb{R}$.

**Theorem C.1.1.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-polynomial continuous function and $a_1, \ldots, a_n \in \mathbb{R}$. Let $\mathcal{P}$ be a partition of $[n]$ and define*

$$\Psi = \{\mathcal{Q} \leq \mathcal{P} \,|\, \sum_{i \in P} a_i = 0 \; \forall P \in \mathcal{Q}\}.$$

*The set $B$ of elements $b = (b_1, \ldots, b_n) \in \mathbb{R}^{n \times m}$ which satisfy*

$$\sum_{P \in \mathcal{P}} \sum_{i \in P} a_i \sigma\Big(b_i \cdot x + y_P\Big) = 0 \qquad \Big(\forall x \in \mathbb{R}^m \, \forall y = (y_P)_{P \in \mathcal{P}} \in \mathbb{R}^{\mathcal{P}}\Big) \qquad \text{(C.1)}$$

*is*

$$\bigcup_{\mathcal{Q}\in\Psi} \{(b_1,\dots,b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall\{i_1,\dots,i_k\} \in \mathcal{Q}\}. \tag{C.2}$$

*Equivalently,*

$$B = \bigcup_{\mathcal{Q}\in\Psi} \bigcap_{\substack{P\in\mathcal{Q}\\ i,j\in P}} \{(b_1,\dots,b_n) \mid b_i - b_j = 0\}.$$

*For arbitrary continuous functions $\sigma$, it is only true that the set defined in* (C.2) *is contained in $B$.*

To prove Theorem C.1.1, we first need to prove some auxiliary results. Theorem C.1.2, stated below, is a reformulation of Theorem 2.27 in [62] adapted here to the context of continuous real functions for convenience. For further discussion, refer to Appendix C.2.

**Theorem C.1.2.** *Let $a_1,\dots,a_n$ non-null real values, and let $b_1,\dots,b_n \in \mathbb{R}^m$ be distinct real vectors. Continuous solutions $\sigma : \mathbb{R} \to \mathbb{R}$ of*

$$\sum_i a_i \sigma\Big(b_i \cdot x + y\Big) = 0 \qquad (\forall x \in \mathbb{R}^m \, \forall y \in \mathbb{R}) \tag{C.3}$$

*are polynomial.*

Moreover, to prove Theorem C.1.1, the following notions and auxiliary results are required.

**Definition C.1.3.** Let $b = (b_1,\dots,b_n) \in \mathbb{R}^{n\times m}$, the *identity pattern* of $b$ is the coarser partition $\mathcal{P}$ of $[n]$ such that $b_i = b_j$ for each $i,j \in P$ and $P \in \mathcal{P}$.

**Theorem C.1.4.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-polynomial continuous function, and $a_1,\dots,a_n \in \mathbb{R}$. Then $b = (b_1,\dots,b_n) \in \mathbb{R}^{n\times m}$ satisfies*

$$\sum_{i=1}^{n} a_i \sigma\Big(b_i \cdot x + y\Big) = 0 \qquad (\forall x \in \mathbb{R}^m \, \forall y \in \mathbb{R}) \tag{C.4}$$

*if and only if $\sum_{i\in P} a_i = 0$ for each $P$ in the identity pattern of $b$.*

*Proof.* Let $P_1,\dots,P_q$ be the parts in the identity pattern of $b$ such that $\sum_{i\in P_j} a_i \neq 0$, define $a'_j = \sum_{i\in P_j} a_i$ then we can rewrite the equation in (C.4) as

$$\sum_{j=1}^{q} a'_j \sigma\Big(b'_j \cdot x + y\Big) = 0,$$

where, for each $j = 1,\dots,q$, the value of $b'_j$ is set to the value of the $b_i$s for $i \in P_j$, which are all equal to each other. Since the $a'_j$ are non-null and $b'_j$ are distinct, by Theorem C.1.2, $\sigma$ have to be polynomial which is impossible. To prove the opposite implication, let $\mathcal{P}$ be the identity pattern of $b$ and write

$$\sum_{i=1}^{n} a_i \sigma\Big(b_i \cdot x + y\Big) = \sum_{P\in\mathcal{P}} \sum_{i\in P} a_i \sigma\Big(b_i \cdot x + y\Big) = \sum_{P\in\mathcal{P}} \Big(\sum_{i\in P} a_i\Big) \sigma\Big(b_i \cdot x + y\Big),$$

where the last equality is possible because $b_i = b_j$ for each $i,j \in P$. □

*Remark* C.1.5. Note that the second implication of Theorem C.1.4 holds for any $\sigma$, including polynomial functions.

This theorem gives the following corollary, which is the one actually needed to prove Theorem C.1.1.

**Corollary C.1.6.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-polynomial continuous function and $a_1, \dots, a_n \in \mathbb{R}$. Let $\mathcal{P}_n$ be the set of all partition of $[n]$ and define*

$$\Phi = \{\mathcal{P} \in \mathcal{P}_n \mid \sum_{i \in P} a_i = 0 \ \forall P \in \mathcal{P}\}.$$

*The set $B$ of elements $b = (b_1, \dots, b_n) \in \mathbb{R}^{n \times m}$ satisfying* (C.4) *is*

$$\bigcup_{\mathcal{P} \in \Phi} \{(b_1, \dots, b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall \{i_1, \dots, i_k\} \in \mathcal{P}\}. \tag{C.5}$$

*Or equivalently,*

$$B = \bigcup_{\mathcal{P} \in \Phi} \bigcap_{\substack{P \in \mathcal{P} \\ i,j \in P}} \{(b_1, \dots, b_n) \mid b_i - b_j = 0\}.$$

*For arbitrary continuous functions $\sigma$, it is only true that the set defined in* (C.5) *is contained in $B$.*

*Proof.* Define

$$B' = \bigcup_{\mathcal{P} \in \Phi} \{(b_1, \dots, b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall \{i_1, \dots, i_k\} \in \mathcal{P}\}.$$

By Theorem C.1.4, $b$ satisfies (C.4) if and only if $\sum_{i \in P} a_i = 0$ for each $P$ in the identity pattern of $b$. Thus, $B \subseteq B'$. To prove the opposite inclusion, note that if $b = (b_1, \dots, b_n) \in B'$ then there exist $\mathcal{P} \in \Phi$ such that $b$ has identity pattern $\mathcal{P}$, then, as $\sum_{i \in P} a_i = 0$ for each $P \in \mathcal{P}$, (C.4) is verified. Finally, note that this implication holds for any $\sigma$ by Remark C.1.5, proving the last claim in Corollary C.1.6. □

*Proof of Theorem C.1.1.* Notice that the problem

$$\sum_{P \in \mathcal{P}} \sum_{i \in P} a_i \sigma\Big(b_i \cdot x + y_P\Big) = 0 \qquad \Big(\forall x \in \mathbb{R}^m \, \forall y = (y_P)_{P \in \mathcal{P}} \in \mathbb{R}^{\mathcal{P}}\Big) \tag{C.6}$$

is equivalent to

$$\sum_{P \in \mathcal{P}} \sum_{i \in P} a_i \sigma\Big(b_i \cdot x + y_P + \hat{y}\Big) = 0 \quad \Big(\forall x \in \mathbb{R}^m \, \forall y \in \mathbb{R}^{\mathcal{P}} \, \forall \hat{y} \in \mathbb{R}\Big)$$

through the change of variables $y_P \mapsto y_P + \hat{y}$ for each $P \in \mathcal{P}$. This problem is in turn equivalent to

$$\sum_{i=1}^{n} a_i \sigma\Big(\hat{b}_i \cdot \hat{x} + \hat{y}\Big) = 0 \qquad \Big(\forall \hat{x} \in \mathbb{R}^m \oplus \mathbb{R}^{\mathcal{P}} \, \forall \hat{y} \in \mathbb{R}\Big) \tag{C.7}$$

due to the following change of variables

$$\hat{b}_i \mapsto \begin{pmatrix} b_i \\ e_P \end{pmatrix} \text{ for } i \in P, \text{ and } \hat{x} \mapsto \begin{pmatrix} x \\ y_P e_P \end{pmatrix},$$

where $\{e_P\}_{P\in\mathcal{P}}$ is the canonical base of $\mathbb{R}^{\mathcal{P}}$. Corollary C.1.6 implies that the solution to (C.7) is

$$\bigcup_{\mathcal{Q}\in\Phi} \{(\hat{b}_1,\ldots,\hat{b}_n) \mid \hat{b}_{i_1} = \cdots = \hat{b}_{i_k} \ \forall\{i_1,\ldots,i_k\} \in \mathcal{Q}\}. \tag{C.8}$$

Note that $\hat{b}_{i_1} = \cdots = \hat{b}_{i_k}$ if and only if $b_{i_1} = \cdots = b_{i_k}$ and $\{i_1,\ldots,i_k\} \subseteq P$ for some $P \in \mathcal{P}$ if and only if $b_{i_1} = \cdots = b_{i_k}$ and $\{i_1,\ldots,i_k\} \in \mathcal{Q}$ for some $\mathcal{Q} \leq \mathcal{P}$.

Recall the definitions

$$\Phi = \{\mathcal{Q} \in \mathcal{P}_n \mid \sum_{i\in P} a_i = 0 \ \forall P \in \mathcal{Q}\} \text{ and } \Psi = \{\mathcal{Q} \leq \mathcal{P} \mid \sum_{i\in P} a_i = 0 \ \forall P \in \mathcal{Q}\}.$$

Noting that $\Psi = \{\mathcal{Q} \in \Phi \mid \mathcal{Q} \leq \mathcal{P}\}$, equation (C.8) implies that the solutions of (C.6) are

$$\bigcup_{\mathcal{Q}\in\Psi} \{(b_1,\ldots,b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall\{i_1,\ldots,i_k\} \in \mathcal{Q}\}.$$

The final claim follows directly from the concluding statement in Corollary C.1.6. □

*Remark* C.1.7. In Theorem C.1.1 the union

$$\bigcup_{\mathcal{Q}\in\Psi} \{(b_1,\ldots,b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall\{i_1,\ldots,i_k\} \in \mathcal{Q}\} \tag{C.9}$$

has redundancies. Indeed, $\{(b_1,\ldots,b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall\{i_1,\ldots,i_k\} \in \mathcal{Q}\}$ is contained in $\{(b_1,\ldots,b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall\{i_1,\ldots,i_k\} \in \mathcal{R}\}$ for each $\mathcal{Q} \leq \mathcal{R}$ finer partitions of $\mathcal{P}$. Hence, the set defined by (C.9) is the same as

$$\bigcup_{\mathcal{Q}\in\Psi'} \{(b_1,\ldots,b_n) \mid b_{i_1} = \cdots = b_{i_k} \ \forall\{i_1,\ldots,i_k\} \in \mathcal{Q}\}$$

where $\Psi'$ is the subset of $\Psi$ containing only the minimal partitions with respect to the refinement ordering.

## C.2 Generalized Polynomials in the Continuous Case

In Appendix C.1, we employ Theorem C.1.2, a reformulation of Theorem 2.27 in [62], adapted here for convenience to the context of continuous real functions. In particular, the original version of this theorem proves that arbitrary complex functions satisfying (C.3) are generalized polynomials, defined as follows.

**Definition C.2.1.** A function $\sigma : \mathbb{C} \to \mathbb{C}$ is a *generalized monomial* function if there exist a symmetric function $F : \mathbb{C}^n \to \mathbb{C}$ additive in each of its variables, such that $\sigma(x) = F(x, \dots, x)$ for each $x \in \mathbb{C}$. A function $\sigma : \mathbb{C} \to \mathbb{C}$ is a *generalized polynomial* if it is a finite sum of generalized monomials, we say that a generalized polynomial $\sigma$ is *real* if $\sigma$ is real and there exists a symmetric function $F : \mathbb{C}^n \to \mathbb{R}$ additive in each of its variables, such that $\sigma(x) = F(x, \dots, x)$ for each $x \in \mathbb{R}$.

Trivially, since complex functions satisfying (C.3) are generalized polynomials, any real solutions are real generalized polynomials.

To conclude the proof of Theorem C.1.2, it remains to show that continuous real generalized polynomials are simply real polynomial functions, as shown by Proposition C.2.2.

**Proposition C.2.2.** *A real continuous generalized polynomial is a real polynomial function.*

*Proof.* The proof of Theorem C.2.2 will be analogous to the proof of the classical proof that any continuous real additive function is linear, see Theorem 1.1 in [58].

First, we show that real generalized monomials are monomial functions on rational numbers.

Indeed, suppose first that $f$ is a real generalized monomial and let $F : \mathbb{R}^n \to \mathbb{R}$ be the symmetric function additive in each variable and such that $f(x) = F(x, \dots, x)$ for each $x \in \mathbb{R}$. Note that for each $r \in \mathbb{N}$,

$$\begin{aligned} F(x_1, \dots, rx_i, \dots, x_n) = F(x_1, \dots, x_i + \cdots + x_i, \dots, x_n) = \\ F(x_1, \dots, x_i, \dots, x_n) + \cdots + F(x_1, \dots, x_i, \dots, x_n) = rF(x_1, \dots, x_i, \dots, x_n). \end{aligned} \tag{C.10}$$

Note that $F(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) = 0$, indeed

$$\begin{aligned} F(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) = \\ F(x_1, \dots, x_{i-1}, 0 + 0, x_{i+1}, \dots, x_n) = \\ F(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) + F(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) \end{aligned} \tag{C.11}$$

Eliminating a term $F(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n)$ from both the sides of (C.11), we get the required result.

Furthermore, $F(x_1, \dots, x_i, \dots, x_n) = -F(x_1, \dots, -x_i, \dots, x_n)$. Indeed,

$$F(x_1, \dots, x_i, \dots, x_n) + F(x_1, \dots, -x_i, \dots, x_n) = F(x_1, \dots, x_i - x_i, \dots, x_n) = 0. \tag{C.12}$$

Equations (C.10) and (C.12) yields

$$F(x_1, \dots, rx_i, \dots, x_n) = rF(x_1, \dots, x_i, \dots, x_n) \tag{C.13}$$

for each $r \in \mathbb{Z}$. Note that by substituting $rx_i = y_i$, we obtain

$$F(x_1, \dots, y_i, \dots, x_n) = rF(x_1, \dots, \frac{1}{r}y_i, \dots, x_n)$$

Equivalently,

$$F(x_1, \ldots, \frac{1}{r} y_i, \ldots, x_n) = \frac{1}{r} F(x_1, \ldots, y_i, \ldots, x_n) \tag{C.14}$$

Equations (C.13) and (C.14) prove

$$F(x, \ldots, rx, \ldots, x) = rF(x, \ldots, x) \tag{C.15}$$

for each $r \in \mathbb{Q}$. Hence,

$$f(rx) = F(rx, \ldots, rx) = r^n F(x, \ldots, x) = r^n f(x).$$

for each $r \in \mathbb{Q}$. In particular set $x = 1$ and $f(1) = c \in \mathbb{R}$,

$$f(r) = r^n f(1) = cr^n.$$

Hence, a real generalized monomial is a monomial on $\mathbb{Q}$.

Finally, we can prove the general case where $f$ is a real generalized polynomial. Recalling that real generalized polynomials are sums of real generalized monomials, they are sums of real monomial functions on $\mathbb{Q}$, namely polynomial functions on $\mathbb{Q}$.

We conclude by noting that, since $f$ is continuous, it extends as a polynomial function on $\mathbb{R}$ due to continuity. □

# Appendix D

# Proofs

## D.1 Equivariant Activations

Thanks to these observations we can prove Corollary 3.4.2

*Proof.* Let us study $\mathrm{SO}(3) \times S_n$-equivariant linear layers. Irreducible representations of $\mathrm{SO}(3) \times S_n$ are the tensor products of irreducible representations of $\mathrm{SO}(3)$ and $S_n$ as shown above. The natural action of $\mathrm{SO}(3)$ on $\mathbb{R}^3$ is an irreducible representation and admissible irreducible representations are invariant, i.e., the trivial action of $\mathrm{SO}(3)$ on $\mathbb{R}$. Hence, the fist layer of an $\mathrm{SO}(3) \times S_n$-equivariant network processing a geometric graph will be a linear layer from a direct sum of $\mathbb{R}^3 \otimes S^\lambda$ to a direct sum of $\mathbb{R} \otimes S^\mu$, where $S^\lambda$ and $S^\mu$ indicate irreducible representations for $S_n$. But, due to Schur's Lemma, the only possible linear layer would be the null one, i.e., a layer without trainable parameters. □

### D.1.1 Proof of the Stabilization Lemma

For convenience we restate Lemma 3.1.2.

**Lemma D.1.1.** *The group of matrices $\mathcal{M}' = \mathcal{M}(\mathcal{F}(\mathcal{M}))$ is the largest group in $\mathrm{GL}_n(\mathbb{R})$ for which $\mathcal{F}(\mathcal{M}') = \mathcal{F}(\mathcal{M})$, and $\mathcal{F}' = \mathcal{F}(\mathcal{M}(\mathcal{F}))$ is the largest family of functions in $\mathcal{C}(\mathbb{R})$ for which $\mathcal{M}(\mathcal{F}') = \mathcal{M}(\mathcal{F})$.*

In what follows we state and prove some results necessary for the proof of Lemma 3.1.2. The proof of the next lemma is trivial.

**Lemma D.1.2.** *The two following statements are true.*

1. *For each group of matrices $\mathcal{M}_1$, $\mathcal{M}_1 \subseteq \mathcal{M}(\mathcal{F}(\mathcal{M}_1))$ and for each family of functions $\mathcal{F}_1$, $\mathcal{F}_1 \subseteq \mathcal{F}(\mathcal{M}(\mathcal{F}_1))$,*
2. *For each inclusion $\mathcal{M}_1 \subseteq \mathcal{M}_2$ of groups of matrices, $\mathcal{F}(\mathcal{M}_2) \subseteq \mathcal{F}(\mathcal{M}_1)$, similarly, for families of activations $\mathcal{F}_1 \subseteq \mathcal{F}_2$, $\mathcal{M}(\mathcal{F}_2) \subseteq \mathcal{M}(\mathcal{F}_1)$.*

**Lemma D.1.3.** *For each family of functions $\mathcal{F}_1$, $\mathcal{M}(\mathcal{F}_1) = \mathcal{M}(\mathcal{F}(\mathcal{M}(\mathcal{F}_1)))$. For each group of matrices $\mathcal{M}_1$, $\mathcal{F}(\mathcal{M}_1) = \mathcal{F}(\mathcal{M}(\mathcal{F}(\mathcal{M}_1)))$.*

*Proof.* We prove the equality $\mathcal{M}_1$, $\mathcal{M}(\mathcal{F}_1) = \mathcal{M}(\mathcal{F}(\mathcal{M}(\mathcal{F}_1)))$ first by proving $\mathcal{M}_1$, $\mathcal{M}(\mathcal{F}_1) \subseteq \mathcal{M}(\mathcal{F}(\mathcal{M}(\mathcal{F}_1)))$, then we prove the opposite inclusion.

Substituting $\mathcal{M}_1$ with $\mathcal{M}(\mathcal{F}_1)$ in the first point of Lemma D.1.2, we get $\mathcal{M}(\mathcal{F}_1) \subseteq \mathcal{M}(\mathcal{F}(\mathcal{M}(\mathcal{F}_1)))$.

To prove the opposite inclusion, substitute $\mathcal{F}_2 = \mathcal{F}(\mathcal{M}(\mathcal{F}_1))$ in the second point of Lemma D.1.2, we obtain $\mathcal{M}(\mathcal{F}(\mathcal{M}(\mathcal{F}_1))) \subseteq \mathcal{M}(\mathcal{F}_1)$. Implying $\mathcal{M}(\mathcal{F}_1) = \mathcal{M}(\mathcal{F}(\mathcal{M}(\mathcal{F}_1)))$. In a similar way, we can prove the other equality. □

We are now ready to prove Lemma 3.1.2.

*Proof.* We only prove the first equality as the second have an analogous proof. By Lemma D.1.3, we know that $\mathcal{F}(\mathcal{M}') = \mathcal{F}(\mathcal{M})$. Then, for each $\mathcal{S} \subseteq \mathrm{GL}_n(\mathbb{R})$ such that $\mathcal{F}(\mathcal{S}) = \mathcal{F}(\mathcal{M})$, we know that $\mathcal{S} \subseteq \mathcal{M}(\mathcal{F}(\mathcal{S}))$ by Lemma D.1.2. Moreover, $\mathcal{S} \subseteq \mathcal{M}(\mathcal{F}(\mathcal{S})) = \mathcal{M}(\mathcal{F}(\mathcal{M})) = \mathcal{M}'$. Hence each $\mathcal{S} \subseteq \mathrm{GL}_n(\mathbb{R})$ such that $\mathcal{F}(\mathcal{S}) = \mathcal{F}(\mathcal{M})$ we have $S \subseteq \mathcal{M}'$, therefore $\mathcal{M}' = \mathcal{M}(\mathcal{F}(\mathcal{M}))$ is the largest group in $\mathrm{GL}_n(\mathbb{R})$ such that $\mathcal{F}(\mathcal{M}') = \mathcal{F}(\mathcal{M})$. □

### D.1.2 Proof of the Main Theorem

We can now state our main result, which shows that, under mild assumptions, there are only a very limited number of maximal groups (or equivalently, of maximal function classes). Moreover, for each of these we provide the exact correspondence between $\mathcal{M}$ and $\mathcal{F}$, thus providing an exhaustive classification of all possible admissible pairs of $\mathcal{M}$ and $\mathcal{F}$.

A class of functions fundamental for the understating of what follows will be $T$-equivariant functions which we define as follows.

**Definition D.1.4.** Let $T$ be a multiplicative subgroup of $\mathbb{R}$. We say that a continuous function $f$ is $T$-equivariant if $f(tx) = tf(x)$ for each $t \in T$ and $x \in \mathbb{R}$. We write $\mathcal{F}_T$ to indicate the set of all continuous $T$-equivariant functions.

Note that if $T = \langle b^n \rangle_{n \in \mathbb{Z}}$, the notion of $T$-equivariance reduces to the notion of $\langle b \rangle$-equivariance provided in Section 3.1. Similarly, for $T = \langle \pm b^n \rangle_{n \in \mathbb{Z}}$ and $\langle \pm b \rangle$-equivariance.

We are now ready to state Theorem D.1.5, one of the two key results fundamental to prove Theorem 3.1.3. In particular, Theorem D.1.5 characterizes admissible pairs indexing them as the multiplicative subgroups of $\mathbb{R}$ but does not provide a constructive description of their families of activation functions, this description is given by Proposition 2.2.5. In what follows we will write $\mathcal{R}_n$ for the set of all unit row invertible matrices.

**Theorem D.1.5.** *Maximal admissible pairs for $T = \mathcal{T}(\mathcal{M})$ are*

- $(\mathcal{M}, \mathrm{Aff}(\mathbb{R}, \mathbb{R}))$ *for each group of non-monomial matrices* $\mathcal{M}$ *in* $\mathcal{R}_n(\mathbb{R})$ *and* $(\mathrm{GL}_n(\mathbb{R}), \mathrm{Hom}(\mathbb{R}, \mathbb{R}))$ *if* $T$ *is dense,*
- $(\mathcal{M}_n(T), \mathcal{F}_T)$ *otherwise.*

Let us denote $\mathbb{R}^*$ as the set of non-zero reals and note that it is a multiplicative group. In the following, we will use the following characterization of the multiplicative subgroups of $\mathbb{R}^*$ that we need to prove Proposition 2.2.5.

To prove Theorem D.1.5 we need to state and prove the two following lemmas. The main ideas behind their proofs are primarily due to [121].

**Lemma D.1.6.** *Define the multiplicative group* $\mathcal{T}(\mathcal{M}) = \langle \sum_{j \in S} M_{ij} : S \subseteq [n], M \in \mathcal{M}, \in [n] \rangle \backslash \{0\}$ *and* $\tilde{f}_0(x) = \tilde{f}(x) - \tilde{f}(0)$*. Note that* $\tilde{f}$ *is* $\mathcal{M}$*-equivariant if and only if* $\tilde{f}_0$ *is* $\mathcal{M}$*-equivariant and* $\tilde{f}_0(0)$ *is a* $\mathcal{M}$*-invariant vector. In particular, if* $\tilde{f}$ *is* $\mathcal{M}$*-equivariant then* $f_0$ *is* $\mathcal{T}(\mathcal{M})$*-equivariant and* $f(0) = 0$ *or* $\mathcal{M} \subseteq \mathcal{R}_n$*.*

*Proof.* Note that $\tilde{f}(Mx) = M\tilde{f}(x)$ for each $x \in \mathbb{R}^n$ if and only if $\tilde{f}_0(Mx) = \tilde{f}(Mx) - \tilde{f}(M0) = M\tilde{f}(x) - M\tilde{f}(0) = M\tilde{f}_0(x)$ for each $x \in \mathbb{R}^n$ if and only if $\tilde{f}_0(Mx) = M\tilde{f}_0(x)$ for each $x \in \mathbb{R}^n$ and $M\tilde{f}_0(0) = \tilde{f}_0(0)$. In particular, $M\tilde{f}(0) = \tilde{f}(0)$ if and only if $M\mathbb{1}f(0) = \mathbb{1}f(0)$, where $\mathbb{1}$ is the vector with all ones, if and only if $f(0) = 0$ or $M \in \mathcal{R}_n$. For each $S \subseteq [n]$ and each $x \in \mathbb{R}$ we have that $f_0(\sum_{j \in S} M_{rj} x) = \sum_{j \in S} M_{rj} f_0(x)$ for each $r \in [n]$, hence $f_0$ is $\mathcal{T}(\mathcal{M})$-equivariant as $f_0(M_i x) = M_i f_0(x)$ for each $x \in \mathbb{R}$ and $i = 1, 2$ then $f_0(M_1 M_2 x) = M_1 M_2 f_0(x)$. □

**Lemma D.1.7.** *Let* $T = \mathcal{T}(\mathcal{M})$ *be a non-dense subset of* $\mathbb{R}^*$*. Then* $\mathcal{F}(\mathcal{M})$ *contains non-affine functions if and only if matrices in* $\mathcal{M}$ *are* $T$*-monomial.*

*Proof.* ($\Rightarrow$) For each $f \in \mathcal{F}(\mathcal{M})$, $f_0 = f - f(0)$ is $T$-equivariant and $\tilde{f}_0$ is $\mathcal{M}$-equivariant by Lemma D.1.6. Suppose $\mathcal{M}$ contains a matrix $M$ which fails to be $T$-monomial, without loss of generality we may assume $M_{11} = t_1$ and $M_{12} = t_2$ not zero. Hence, $f_0$ is additive. Indeed, for each $x_1, x_2 \in \mathbb{R}$, we have that

$$f_0(x_1 + x_2) = \langle \tilde{f}_0(M(t_1^{-1} x_1 e_1 + t_2^{-1} x_2 e_2)), e_1 \rangle =$$

$$\langle M\tilde{f}_0(t_1^{-1} x_1 e_1 + t_2^{-1} x_2 e_2), e_1 \rangle = t_1 f_0(t_1^{-1} x_1) + t_2 f_0(t_2^{-1} x_2) = f_0(x_1) + f_0(x_2)$$

Therefore $f_0$ is linear, being both additive and continuous, this implies $f$ to be affine which contradicts the hypothesis.
($\Leftarrow$) If $\mathcal{M}$ contains only $T$-monomial matrices, it is easy to check that each $T$-equivariant function induces an equivariant activation, which are not all affine by Lemma 2.2.8. □

Now we are ready to present the proof of Theorem D.1.5. For convenience in what follows we will write as $\mathcal{P}_n$, the group of $n \times n$ permutation matrices.

*Proof.* We will study which are the groups of matrices $\mathcal{M}$ such that $\mathcal{T}(\mathcal{M}) = T$ as $T$ varies between subgroups of $\mathbb{R}^*$. If $T$ is a dense subgroup of $\mathbb{R}^*$, Lemma 2.2.8 and Lemma D.1.6 implies $\mathcal{F}(\mathcal{M}) = \mathrm{Aff}(\mathbb{R}, \mathbb{R})$ if $\mathcal{M} \subseteq \mathcal{R}_n$ or

$\mathcal{F}(\mathcal{M}) = \mathrm{Hom}(\mathbb{R},\mathbb{R})$ otherwise. In those cases, $\mathcal{M}(\mathrm{Aff}(\mathbb{R},\mathbb{R})) \subseteq \mathcal{R}_n(\mathbb{R})$ and $\mathcal{M}(\mathrm{Hom}(\mathbb{R},\mathbb{R})) = \mathrm{GL}_n(\mathbb{R})$. For dense $T$, Lemma 3.1.2 implies that admissible maximal pairs are $(\mathcal{M} \subseteq \mathcal{R}_n, \mathrm{Aff}(\mathbb{R},\mathbb{R}))$ and $(\mathrm{GL}_n(\mathbb{R}), \mathrm{Hom}(\mathbb{R},\mathbb{R}))$, whose family of activations only contains affine functions.

If $T$ is non-dense and non-trivial, by Lemma D.1.7 we have two cases: if $\mathcal{F}(\mathcal{M})$ only contains affine functions we reduce to the maximal admissible pairs $(\mathcal{M} \subseteq \mathcal{R}_n(\mathbb{R}), \mathrm{Aff}(\mathbb{R},\mathbb{R}))$ and $(\mathrm{GL}_n(\mathbb{R}), \mathrm{Hom}(\mathbb{R},\mathbb{R}))$, otherwise, due to Lemma D.1.7, $\mathcal{M} < \mathcal{M}_n(T)$ and $\mathcal{F}(\mathcal{M}) = \mathcal{F}_T$ by Lemma D.1.6 and $\mathcal{M}(\mathcal{F}_T) = \mathcal{M}_n(T)$ because the inclusion $\mathcal{M}_n(T) \subseteq \mathcal{M}(\mathcal{F}_T)$ is obvious and if $\mathcal{M}(\mathcal{F}_T)$ contains non-monomial matrices, Lemma D.1.7 would contradict $\mathcal{F}_T$ containing non-affine functions.

Applying Lemma 3.1.2, we get the admissible maximal pairs $(\mathcal{M}_n(T), \mathcal{F}_T)$. Finally, if $T$ is trivial, $\mathcal{M} = \mathcal{P}_n$ hence we obtain the maximal admissible pair $(\mathcal{P}_n, \mathcal{C}(\mathbb{R}))$ as verifying $\mathcal{F}(\mathcal{P}_n) = \mathcal{C}(\mathbb{R})$ is trivial and Lemma D.1.7 implies $\mathcal{M}(\mathcal{C}(\mathbb{R})) = \mathcal{M}_n(\{1\}) = \mathcal{P}_n$. □

We are now ready to prove Theorem 3.1.3.

*Proof.* Theorem D.1.5 classifies admissible pairs whose relevant families of activations are $\mathcal{F}_T$ where $T$ varies on the multiplicative subgroups of $\mathbb{R}$. Thanks to the complete classification of $T$-equivariant functions, $\mathcal{F}_T$, provided by Proposition 2.2.5, we can substitute them with their concrete counterparts enumerated in the statement. □

In what follows we restate Theorem 3.1.4 and prove it.

**Theorem D.1.8.** *Let $G$ be a compact group and let us restrict to families of activations containing some non-affine functions. The only two maximal admissible pairs up to isomorphism of groups of matrices are*

- *Continuous functions and permutation matrices,*
- *Odd continuous functions and signed permutation matrices.*

*Proof.* At first, we want to prove that each admissible representation of $G$ is isomorphic either to a permutation representation or to a sign-permutation representation. We split the proof into two parts. In the first part, $G$ is represented by non-negative monomial matrices and, in the second, $G$ is represented by arbitrary monomial matrices. Thanks to Theorem 3.1.3, those two cases covers all the admissible cases for families of activations containing at least one non-affine function.

By [36], each bounded non-negative group of matrices is isomorphic to a permutation group of matrices by a positive scaling of basis. Hence, the image of a compact group in $\mathcal{M}_n(T)$ with $T \subseteq \mathbb{R}_{>0}$ can be written as a group of permutation matrices after a positive scaling of basis. Hence, all the considered group of matrices are isomorphic to $\mathcal{P}_n$, therefore reducing to the pair $(\mathcal{P}_n, \mathcal{C}(\mathbb{R}))$.

Now consider $T$ an arbitrary non-dense multiplicative subgroup of $\mathbb{R}$. Each monomial matrix can be written as the product $SDP$, where $S$ is a diagonal

matrix containing only $\pm 1$, $D$ a positive diagonal matrix, and $P$ is a permutation matrix. Consider the map $\phi : SDP \mapsto DP$. Note that $\phi$ is a continuous group homomorphism and that its image is a compact group of non-negative matrices. Indeed, $\phi$ is just the absolute value function defined on all the elements of the matrices, hence it is continuous. Then, let $S_1 D_1 P_\sigma$ and $S_2 D_2 P_\tau$ be two monomial matrices as before, where $P_\sigma$ and $P_\tau$ are permutation matrices respectively representing permutations $\sigma$ and $\tau$. Then their composition $S_1 D_1 P_\sigma S_2 D_2 P_\tau = S_1 \sigma(S_2) D_1 P_\sigma D_2 P_\tau$ where $\sigma(S_2)$ is the diagonal matrix obtained by permuting the diagonal elements through $\sigma$, i.e. $\sigma(S_2) = P_\sigma^t S_2 P_\sigma$, which commutes with $D_1$ being both diagonal matrices. This proves that $\phi$ is an homomorphism.

For what we proved at the beginning of the proof, there exists a positive scaling, represented by a positive diagonal matrix $B$, such that $BDPB^{-1} = P'$ is a permutation matrix for each $DP$ in the image of $\phi$. Note that for each $SDP$, after scaling by $B$, we obtain $BSDPB^{-1} = SBDPB^{-1} = SP'$ a signed permutation matrix. This is true because $S$ and $B$ commute being both diagonal matrices. Hence, all the considered groups of matrices are isomorphic to the group of signed permutation matrices, therefore, therefore reducing to the pair of odd continuous functions and signed permutation matrices.

Now that we have proven that there are only two isomorphism classes of admissible representations for $G$, we want to show that there is only one representation in each class with maximal family of activation functions. Let $\mathcal{M}$ be the image of a positive monomial matrices representation of $G$. By Lemma D.1.6, each admissible group of matrices $\mathcal{M}'$ isomorphic to $\mathcal{M}$ commutes with $\mathcal{F}(\mathcal{M}') = \mathcal{F}_T$ for some $T$. The maximal family $\mathcal{F}_T$ is $\mathcal{C}(\mathbb{R})$ and it happens when $T = \langle 1 \rangle$ and $\mathcal{M}'$ are permutation matrices. Note that the isomorphism between $\mathcal{M}$ and some group of permutation matrices is always possible as shown at the beginning of this proof. An analogous argument applies to groups of arbitrary monomial matrices.

We conclude by noticing that in general permutation representations and signed-permutation representations are not isomorphic, e.g. the trivial representation and the sign representation of $S_n$. This proves that the two presented pairs are *in general* disjoint. □

Note that compactness is a required condition. Indeed, given $b > 1$, consider the following one-dimensional representation representation $\rho : \mathbb{Z} \to \mathrm{GL}_1(\mathbb{R}) = \mathbb{R}^*$ such that $\rho(n)x = b^n x$. The image $\rho(\mathbb{Z})$ is not bounded and hence is not compact. This means that $\rho$ cannot be isomorphic to a permutation representation whose image is compact.

## D.2 Separation Constraints

### D.2.1 The Role of Depth

**Lemma D.2.1.** *Let $\mathcal{C} = \{C_1, \dots, C_d\}$ be a finite collection of sets. The following statements are true:*

- *Let* $\mathcal{C}_\cup = \{C_{i_1} \cup \cdots \cup C_{i_r} \mid 1 \leq i_1, \ldots, i_r \leq d,\ r \in \mathbb{N}\}$ *be the collection of unions of a finite number of sets in* $\mathcal{C}$*. Then* $\mathcal{C}_\cup$ *is finite.*
- *Let* $\mathcal{C}_\cap = \{C_{i_1} \cap \cdots \cap C_{i_r} \mid 1 \leq i_1, \ldots, i_r \leq d,\ r \in \mathbb{N}\}$ *be the collection of intersections of a finite number of sets in* $\mathcal{C}$*. Then* $\mathcal{C}_\cap$ *is finite.*
- *Let* $\tilde{\mathcal{C}}$ *the smaller collection containing* $\mathcal{C}$ *which is closed by intersection and union. Then* $\tilde{\mathcal{C}} = (\mathcal{C}_\cap)_\cup = (\mathcal{C}_\cup)_\cap$ *and, in particular, is finite.*

*In particular, ascending and descending sequences of inclusions in* $\tilde{\mathcal{C}}$ *stabilize.*

*Proof.* To prove the first point, it is sufficient to note that duplicates in the expression $C_{i_1} \cup \cdots \cup C_{i_r}$ can be removed. Therefore, the cardinality of $\mathcal{C}_\cup$ is bounded by the number of possible tuples $i_1, \ldots, i_r$ which are $2^d$. The proof of the second point is analogous.

By the distributive property of intersections with respect to unions we obtain that each element in $\tilde{\mathcal{C}}$ can be written as

$$(C_{i_{1,1}} \cap \cdots \cap C_{i_{1,d_1}}) \cup \cdots \cup (C_{i_{r,1}} \cap \cdots \cap C_{i_{r,d_r}}).$$

Hence, $\tilde{\mathcal{C}} = (\mathcal{C}_\cap)_\cup$. Similarly, using the distributive property of unions with respect to intersections, we get $\tilde{\mathcal{C}} = (\mathcal{C}_\cup)_\cap$. In particular, $\tilde{\mathcal{C}}$ is finite as $\mathcal{C}_\cup$ and, hence, $(\mathcal{C}_\cup)_\cap$ are finite. □

The repetition threshold may vary depending on the model and representation. For example, $k$-IGNs, being equivalent to $k$-WL, have a repetition threshold proportional to that of $k$-WL itself [77, 41]. In contrast, the Proposition 5.2.7 demonstrates an example of stabilization after just one repetition.

*Proof of Proposition 5.2.7.* From previous observations, we know that

$$\rho(\mathcal{N}_\sigma(V, \mathbb{R}^G, \ldots, \mathbb{R}^G, \mathbb{R}^{G/H})) \subseteq \rho(\mathcal{N}_\sigma(V, \mathbb{R}^G, \mathbb{R}^{G/H})). \tag{D.1}$$

Note that the family of equivariant continuous functions $\mathcal{C}_G(V, \mathbb{R}^{G/H})$ cannot separate $H$-orbits in $V$. Indeed, for each $f \in \mathcal{C}_G(V, \mathbb{R}^{G/H})$, $f(hv) = hf(v) = f(v)$ for each $h \in H$. Hence, Proposition 5.2.12 implies that $\mathcal{N}_\sigma(V, \mathbb{R}^G, \mathbb{R}^{G/H})$ has the finer separation power between families of functions in $\mathcal{C}_G(V, \mathbb{R}^{G/H})$. This implies equality in (D.1), concluding the proof. □

### D.2.2 The Role of Intermediate Representations

*Proof of Theorem 5.2.8.* For space constraints, given a map $\phi$, we write its parallelization $\Pi(\phi)$ as $\overline{\phi}$. Note that

$$\overline{\psi \circ \phi} = \overline{\psi} \circ \overline{\phi}.$$

Indeed, $\overline{\psi} \circ \overline{\phi} = (\phi, \phi) \circ (\psi, \phi) = (\phi \circ \psi, \phi \circ \psi) = \overline{\phi \circ \psi}$, and $\tilde{\sigma}\iota' = \iota'\tilde{\sigma}$ and $\tilde{\sigma}\pi' = \pi'\tilde{\sigma}$. Similarly, for $\pi''$ and $\iota''$.

Furthermore, $\overline{(\iota' + \iota'') \circ (\pi' + \pi'')} = (\overline{\iota'} + \overline{\iota''}) \circ (\overline{\pi'} + \overline{\pi''})$ since

$$\begin{aligned}\overline{(\iota' + \iota'') \circ (\pi' + \pi'')} &= \overline{id_{V_i' \oplus V_i''}} = \left(\overline{id_{V_i'} \oplus 0_{V_i''}}\right) + \left(\overline{0_{V_i'} \oplus id_{V_i''}}\right) \\ &= \left(\overline{\iota' \circ \pi'}\right) + \left(\overline{\iota'' \circ \pi''}\right) = \left(\overline{\iota'} \circ \overline{\pi'}\right) + \left(\overline{\iota''} \circ \overline{\pi''}\right) \\ &= \left(\overline{\iota'} + \overline{\iota''}\right) \circ \left(\overline{\pi'} + \overline{\pi''}\right).\end{aligned}$$

We now need to prove that

$$\overline{(\psi\iota' + \psi\iota'')}\,\tilde{\sigma}\,\overline{(\pi'\phi + \pi''\phi)} = (\overline{\psi\iota'} + \overline{\psi\iota''})\,\tilde{\sigma}\,(\overline{\pi'\phi} + \overline{\pi''\phi}). \tag{D.2}$$

Indeed,

$$\begin{aligned}\overline{(\psi\iota' + \psi\iota'')}\,\tilde{\sigma}\,\overline{(\pi'\phi + \pi''\phi)} &= \overline{\psi} \circ \overline{(\iota' + \iota'')}\tilde{\sigma}\overline{(\pi' + \pi'')} \circ \overline{\phi} \\ &= \overline{\psi} \circ \overline{(\iota' + \iota'')(\pi' + \pi'')} \circ \tilde{\sigma}\overline{\phi} = \overline{\psi} \circ (\overline{\iota'} + \overline{\iota''})(\overline{\pi'} + \overline{\pi''}) \circ \tilde{\sigma}\overline{\phi} \\ &= \overline{\psi} \circ (\overline{\iota'} + \overline{\iota''})\tilde{\sigma}(\overline{\pi'} + \overline{\pi''})\overline{\phi} = (\overline{\psi\iota'} + \overline{\psi\iota''})\,\tilde{\sigma}\,(\overline{\pi'\phi} + \overline{\pi''\phi}).\end{aligned}$$

Hence, thanks to (D.2),

$$\begin{aligned}&\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{M}_{d-1}, M_d') = \\ &\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i) + \pi_*''(M_i)}, \overline{\iota'^*(M_{i+1}) + \iota''^*(M_{i+1})}, \ldots, \overline{M}_{d-1}, M_d') = \\ &\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)} + \overline{\pi_*''(M_i)}, \overline{\iota'^*(M_{i+1})} + \overline{\iota''^*(M_{i+1})}, \ldots, \overline{M}_{d-1}, M_d').\end{aligned}$$

By Theorem 5.2.2 and the previous observations, we can limit to study spaces of the type

$$\begin{aligned}\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)} + \overline{\pi_*''(M_i)}, (\overline{\iota'^*(M_{i+1})} + \overline{\iota''^*(M_{i+1})})_{uv}) = \\ \mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)} + \overline{\pi_*''(M_i)}, \overline{(\iota'^*(M_{i+1}))}_{uv}) + \\ \mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)} + \overline{\pi_*''(M_i)}, \overline{(\iota''^*(M_{i+1}))}_{uv})\end{aligned}$$

thanks to the linearity of the map $\phi \mapsto (\phi)_{uv}$. Note that

$$\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)} + \overline{\pi_*''(M_i)}, \overline{(\iota'^*(M_{i+1}))}_{uv}) = \mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)}, \overline{(\iota'^*(M_{i+1}))}_{uv})$$

as $\pi' \circ \iota'' = 0$ and both projections and immersions commute with activations. From Lemma D.2.5 we get

$$\begin{aligned}&\mathcal{V}(\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)} + \overline{\pi_*''(M_i)}, (\overline{\iota'^*(M_{i+1})} + \overline{\iota''^*(M_{i+1})})_{uv})) = \\ &\quad = \mathcal{V}(\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*'(M_i)}, \overline{(\iota'^*(M_{i+1}))}_{uv}) \cap \mathcal{V}(\mathcal{N}_\sigma(\overline{M}_1, \ldots, \overline{\pi_*''(M_i)}, \overline{(\iota''^*(M_{i+1}))}_{uv}).\end{aligned}$$

Combining all the above results, we conclude the proof of the theorem. □

### D.2.3 The Role of Representation Type

We are now going to develop the tools to prove Proposition 5.2.12.

**Lemma D.2.2.** *Let $M = \mathrm{Aff}_G(V, \mathbb{R}^G)$, then*

$$\mathcal{V}(M_{u,v}) = \mathcal{V}(M_{v^{-1}u,e}) = \mathcal{V}(M_{uv^{-1},e}).$$

*Moreover, $(\beta, \beta') \in \mathcal{V}(M_{g,e})$ if and only if $g\beta = \beta'$.*

*Proof.* Let $V = V_1 \oplus \cdots \oplus V_s$ where $V_i = \mathbb{R}^{G/K_i}$ for each $i = 1, \ldots, s$. By Proposition 4.1.11 and setting $H = \{e\}$, we know that

$$\mathrm{Hom}_G(V, \mathbb{R}^G) = \mathrm{Hom}_G(\mathbb{R}^{G/K_1}, \mathbb{R}^G) \oplus \cdots \oplus (\mathbb{R}^{G/K_s}, \mathbb{R}^G)$$

is generated by functions $\mathcal{R}_{g_iK_i}\pi_{G/K_i}$ for each $g_iK_i \in G/K_i$ for each $i = 1, \ldots, s$, and $\pi_{G/K_i}$ is the projection of $V$ onto $V_i = \mathbb{R}^{G/K_i}$. Moreover,

$$\Big(\mathcal{R}_{g_iK_i}\pi_{G/K_i}(\beta)\Big)_u = \Big(\mathcal{R}_{g_iK_i}\pi_{G/K_i}\Big(\sum_{kK_i \in G/K_i} \beta_{kK_i} e_{kK_i}\Big)\Big)_u = \frac{1}{|K_i|}\beta_{ug_iK_i}.$$

For each $g \in G$, we have that

$$\begin{aligned}
&\mathcal{V}(M_{u,v}) = \\
&\Big\{(\beta, \beta') \mid \Big(\mathcal{R}_{g_iK_i}\pi_{G/K_i}(\beta)\Big)_u - \Big(\mathcal{R}_{g_iK_i}\pi_{G/K_i}(\beta')\Big)_v = 0 \ \forall i \forall g_iK_i \in G/K_i\Big\} = \\
&\big\{(\beta, \beta') \mid \beta_{ug_iK_i} - \beta'_{vg_iK_i} = 0 \ \forall i \forall g_iK_i \in G/K_i\big\} = \\
&\big\{(\beta, \beta') \mid v^{-1}u\beta = \beta'\big\}.
\end{aligned}$$

In particular, we have that $\mathcal{V}(M_{u,v}) = \mathcal{V}(M_{v^{-1}u,e})$. Hence, $(\beta, \beta') \in \mathcal{V}(M_{g,e})$ if and only if $g\beta = \beta'$.

Finally, in a similar way, we are able to observe that $\mathcal{V}(M_{u,v}) = \mathcal{V}(M_{uv^{-1},e})$. □

*Proof of Proposition 5.2.12.* In what follows we have to consider $G \sqcup G$, to distinguish the two distinct copies of $G$, we denote $G'$ as the second copy of $G$ and, and when $g$ is an element of $G$, we will indicate as $g'$ the analogous element in $G'$.

Define

$$M = \Psi(\mathrm{Aff}_G(V, \mathbb{R}^G))$$

and

$$N = \mathrm{T}(\mathrm{Aff}_G(\mathbb{R}^G, \mathbb{R}^{G/H})) \subseteq \mathrm{Hom}_G(\mathbb{R}^G \oplus \mathbb{R}^{G'}, \mathbb{R}^{G/H}).$$

Proposition 5.2.1 implies

$$\rho(\mathcal{N}_\sigma(V, \mathbb{R}^G, \mathbb{R})) = \mathcal{V}(\mathcal{N}_\sigma(M, N)).$$

Note that $N = \langle \mathcal{R}_{Hg} - \mathcal{R}_{Hg'} \rangle_{Hg \in H\backslash G}$ where functions $\mathcal{R}_{Hg}$ are defined as

$$\Big(\mathcal{R}_{Hg}(e_k)\Big)_{sH} = \begin{cases} 1 & \text{if } s \in kg^{-1}H, \\ 0 & \text{otherwise.} \end{cases}$$

An element $\mathcal{Q}$ in $\Psi_{Hg,sH}$ is a partition of $G \sqcup G'$, where for each $P \in \mathcal{Q}$ the intersection $P \cap sHg \sqcup sH'g$ have the same number of elements in $sHg$ and $sH'g$. Due to Remark C.1.7, we can just consider $\Psi'_{Hg,sH}$ containing the partitions of $G \sqcup G'$ whose only parts are $P = \{u, v\}$ for $u \in sHg$ and $v \in sH'g$, otherwise $P$ is a singleton not containing elements in $sHg$ or $sH'g$.

Hence, by Theorem 5.2.2,

$$\mathcal{V}(\mathcal{N}_\sigma(M, N)) = \bigcap_{Hg,sH} \bigcup_{\mathcal{Q} \in \Psi'_{Hg,sH}} \bigcap_{\{u,v\} \in \mathcal{Q}} \mathcal{V}(M_{u,v}). \tag{D.3}$$

If we prove that for each $Hg$ and $sH$

$$\bigcup_{\mathcal{Q} \in \Psi'_{Hg,sH}} \bigcap_{\{u,v\} \in \mathcal{Q}} \mathcal{V}(M_{u,v}) = \bigcup_{h \in H} \mathcal{V}(M_{h,e}). \tag{D.4}$$

then we are done. Indeed, thanks to Lemma D.2.2, $(\beta, \beta') \in \bigcup_{h \in H} \mathcal{V}(M_{h,e})$ if and only if there exists some $h \in H$ such that $h\beta = \beta'$. Moreover, (D.4) does not depend on $Hg$ and $sH$ then the outer intersection in (D.3) is trivial.

Now to prove (D.4), we first show that

$$\bigcup_{\mathcal{Q} \in \Psi'_{Hg,sH}} \bigcap_{\{u,v\} \in \mathcal{Q}} \mathcal{V}(M_{u,v}) \subseteq \bigcup_{h \in H} \mathcal{V}(M_{h,e}).$$

Note that if $\{u, v\} \in \mathcal{Q}$ and $u, v \in sHg$, then, by Lemma D.2.2,

$$\mathcal{V}(M_{u,v}) = \mathcal{V}(M_{shg,sh'g}) = \mathcal{V}(M_{hh'^{-1},e}).$$

Therefore,

$$\bigcap_{\{u,v\} \in \mathcal{Q}} \mathcal{V}(M_{u,v}) \subseteq \mathcal{V}(M_{u,v}) \subseteq \bigcup_{h \in H} \mathcal{V}(M_{h,e}).$$

The right-hand side is independent of $\mathcal{Q}$ then the union on each $\mathcal{Q}$ in $\Psi'_{Hg,sH}$ of sets on the left-hand side proves the searched inclusion.

To prove the opposite inclusion, for each $h$ define $\mathcal{P}_h \in \Psi'_{Hg,sH}$ as the partition containing the sets $\{ghts, gts\}$ for each $t \in H$ and the remaining singletons. Then, note that, by Lemma D.2.2,

$$\bigcap_{\{ghts,gts\} \in \mathcal{P}_h} \mathcal{V}(M_{ghts,gts}) = \mathcal{V}(M_{h,e}).$$

Hence,

$$\bigcup_{h \in H} \mathcal{V}(M_{h,e}) = \bigcup_{h \in H} \bigcap_{\{ghts,gts\} \in \mathcal{P}_h} \mathcal{V}(M_{ghts,gts}) \subseteq \bigcup_{\mathcal{Q} \in \Psi'_{Hg,sH}} \bigcap_{\{u,v\} \in \mathcal{Q}} \mathcal{V}(M_{u,v}).$$

This concludes the proof.

□

### D.2.4 Implications on Specific Models

**Lemma D.2.3.** *An element $(\alpha, \beta) \in \rho(1\text{-}CNN)$ if and only if there exist a permutation of $\sigma \in S_n$ such that $\alpha_i = \beta_{\sigma(i)}$ for each $i = 1, \dots, n$.*

*Proof.* Write $[n] \sqcup [n]' = \{1, \dots, n, 1', \dots, n'\}$, and notice that $\mathrm{Aff}_{\mathbb{Z}_n}(\mathbb{R}^n, \mathbb{R})' = \langle \mathbb{1}_{[n]} - \mathbb{1}_{[n]'} \rangle$, hence $\Psi'$ as defined in Remark C.1.7 is composed by partitions $\mathcal{Q}$ of $[n] \sqcup [n]'$ such that

$$\mathcal{Q} = \{\{i, j'\} \mid i \in [n], j \in [n]'\}.$$

Recall $C = \langle id_{\mathbb{R}^n \oplus \mathbb{R}^{n'}} \rangle$. Note that $(\alpha, \beta) \in \mathcal{V}(M^1_{i,j'}) = \langle id_{\mathbb{R}^n \oplus \mathbb{R}^{n'}} \rangle$ if and only if $\alpha_i = \beta_j$. Moreover, for a given $\mathcal{Q}$ in $\Psi'$, we have $(\alpha, \beta) \in \bigcap_{i,j' \in \mathcal{Q}} \mathcal{V}(C_{i,j'})$ if and only if, given the bijection $\sigma : [n] \to [n]'$ associating $i$ to $j'$, $\alpha_i = \beta_{\sigma(i)}$ for each $i = 1, \dots, n$.

Notice that, by Theorem 5.2.2,

$$(\alpha, \beta) \in \rho(\mathcal{N}_\sigma(C, \mathrm{Aff}_{\mathbb{Z}_n}(\mathbb{R}^n, \mathbb{R})) = \bigcup_{\mathcal{Q} \in \Psi} \bigcap_{i,j' \in \mathcal{Q}} \mathcal{V}(M^1_{i,j'}),$$

which is equivalent at saying that there exist a permutation of $\sigma \in S_n$ such that $\alpha_i = \beta_{\sigma(i)}$ for each $i = 1, \dots, n$. □

*Proof of Proposition 5.3.3.* Note that $\rho(1\text{-CNN})$ can be characterized by applying Lemma D.2.3 as follows: $(\alpha, \beta) \in \rho(1\text{-CNN})$ if and only if there exists a permutation $\sigma \in S_n$ such that $\alpha_i = \beta_{\sigma(i)}$ for each $i = 1, \dots, n$. In contrast, Proposition 5.2.12 shows that $(\alpha, \beta) \in \rho(\mathcal{N}_\sigma(C^n, \mathrm{Aff}_{\mathbb{Z}_n}(\mathbb{R}^n, \mathbb{R})))$ if and only if there exists an element $g \in \mathbb{Z}_n$ such that $\alpha_i = \beta_{i+g \pmod n}$ for each $i = 1, \dots, n$. Notice that for $n > 1$, $\mathbb{Z}_n \lneq S_n$, hence $\rho(n\text{-CNN}) \subsetneq \rho(1\text{-CNN})$, as desired. The proof of the chain of inclusions

$$\rho(n\text{-CNN}) \subseteq \cdots \subseteq \rho(2\text{-CNN}) \subseteq \rho(1\text{-CNN})$$

is a direct consequence of Lemma D.2.4 since: 1-CNN $\subseteq$ 2-CNN $\subseteq \cdots \subseteq n$-CNN. □

### D.2.5 Technical Lemmas

In what follows, let $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{F}$ be families of functions in $\mathcal{C}(X, V)$, where $X$ is a topological space and $V$ a real vector space.

**Lemma D.2.4.** *If $\mathcal{C} \subseteq \mathcal{D}$, then $\rho(\mathcal{D}) \subseteq \rho(\mathcal{C})$.*

**Lemma D.2.5.** *Let $\mathcal{C}$ and $\mathcal{D}$ be two families of real-valued functions such that each of them contains at least a constant function. The equivalence relations induced by their separation condition are linked by the following conditions $\rho(\mathcal{C} + \mathcal{D}) = \rho(\mathcal{C} \cup \mathcal{D}) = \rho(\mathcal{C}) \cap \rho(\mathcal{D})$.*

*Proof.* Let us prove the first equality. Let $c$ be the constant function in $\mathcal{D}$. Hence $\rho(\mathcal{C}+\mathcal{D}) \subseteq \rho(\mathcal{C}+c) = \rho(\mathcal{C}) \subseteq \rho(\mathcal{C}) \cup \rho(\mathcal{D})$. To prove the inverse inclusion, suppose there exists a function $f$ either in $\mathcal{C}$ or $\mathcal{D}$ separating $x$ and $y$. Without loss of generality, suppose $f \in \mathcal{C}$, $f + c \in \mathcal{C} + \mathcal{D}$ would be separating $x$ and $y$. This conclude the proof of the first equality. The proof of the second equality follows from the definition of $\rho$. Indeed,

$$\begin{aligned}\rho(\mathcal{C} \cup \mathcal{D}) &= \{(x,y) \in X \times X \mid f(x) = f(y)\ \forall f \in \mathcal{C} \cup \mathcal{D}\}. \\ &= \Big\{(x,y) \in X \times X \ \Big|\ \forall f \in \mathcal{C},\ f(x) = f(y)\Big\} \\ &\quad \cap\ \Big\{(x,y) \in X \times X \ \Big|\ \forall f \in \mathcal{D},\ f(x) = f(y)\Big\}. \\ &= \rho(\mathcal{C}) \cap \rho(\mathcal{D}).\end{aligned}$$

□

*Remark* D.2.6. Note that, with slight modifications to the proofs, analogous results to all previous lemmas can be derived by substituting $\rho$ with $\mathcal{V}$.

**Lemma D.2.7.** *If $\mathcal{F}_d$ is a set spanning a null-bias space $M_d$, then*

$$\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_d)) = \mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d)).$$

*Proof.* Trivially,

$$\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d)) \subseteq \mathcal{N}_\sigma(M_1, \dots, M_d)).$$

For the zero-locus analogous of Lemma D.2.4,

$$\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_d)) \subseteq \mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d)).$$

To prove the opposite inclusion, write $\mathcal{F}_d = \{\phi_1, \dots, \phi_s\}$ and note that each neural network $\eta^d$ in $\mathcal{N}_\sigma(M_1, \dots, M_d)$ can be written as

$$\eta^d = (x_1\phi_1 + \cdots + x_s\phi_s) \circ \tilde{\sigma} \circ \eta^{d-1} = x_1(\phi_1 \circ \tilde{\sigma} \circ \eta^{d-1}) + \cdots + x_s(\phi_s \circ \tilde{\sigma} \circ \eta^{d-1}),$$

for some $x_1, \dots, x_s \in \mathbb{R}$ and $\eta^{d-1} \in \mathcal{N}_\sigma(M_1, \dots, M_{d-1})$.

Moreover, note that

$$\phi_i \circ \tilde{\sigma} \circ \eta^{d-1} \in \mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d),$$

for each $i = 1, \dots, s$.

If $\beta \in \mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d))$, then

$$\eta^d(\beta) = x_1(\phi_1 \circ \tilde{\sigma} \circ \eta^{d-1}) + \cdots + x_s(\phi_s \circ \tilde{\sigma} \circ \eta^{d-1}) = 0.$$

Thus,

$$\mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_{d-1}, \mathcal{F}_d)) \subseteq \mathcal{V}(\mathcal{N}_\sigma(M_1, \dots, M_d)),$$

completing the proof. □

## D.3 Shallow Equivariant Networks

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

**Definition D.3.1** (Symmetric Superpositions)**.** Let $\phi_1, \dots, \phi_\ell : \mathbb{R}^n \to \mathbb{R}^d$ be linear maps. We define symmetric superpositions of ridge functions as follows:

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

**Proposition D.3.2.** *The family of functions approximated by $\mathcal{U}_\sigma(M, N)$ coincides with the class $\Delta(\phi_1, \dots, \phi_\ell)$, where $\phi_1, \dots, \phi_\ell$ are the basis maps associate to $M$.*

*Proof of Proposition D.3.2.* In the general setting, write the linear parts of $M$ and $N$ respectively as $\lambda(M) = \operatorname{Span}\left\{\phi^1, \dots, \phi^m\right\}$ and $\lambda(N) = \operatorname{Span}\left\{x \mapsto \mathbb{1}^t \cdot x\right\}$. 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

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

While elements in $N^{h\times 1}$ can be represented as affine maps $x \mapsto Ax + d$ where $d \in \mathbb{R}$ and

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

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

$$\eta(x) = \sum_{j=1}^{h} a_j \sum_{i\in Y} \sigma\left(\sum_{t=1}^{m} b_{j,t}\phi_i^t(x) + c_j\right) = \tag{D.5}$$

$$\sum_{i\in Y}\sum_{j=1}^{h} a_j\sigma\left(\sum_{t=1}^{m} b_{j,t}\phi_i^t(x) + c_j\right) = \sum_{i\in Y} \zeta(\phi_i^1(x), \dots, \phi_i^m(x)) \tag{D.6}$$

where

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

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 (D.5) we can approximate any superposition in $\Delta(\phi_1, \dots, \phi_l)$. Thus, we have

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

On the other hand, by (D.5),

$$\bigcup_{h \in \mathbb{N}} \mathcal{N}_\sigma(M^{1\times h}, N^{h\times 1}) \subseteq \Delta(\phi_1, \dots, \phi_l).$$

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

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

**Proposition D.3.3.** *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 $\phi_i$ are those defined in* (6.1).

*Proof.* By the combination of Proposition D.3.2 and Theorem 8 in [84], 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 $\gamma$ 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 $\gamma$ and zero on all the other values in consideration. In this case,

$$\begin{aligned} F(x) &= f \circ \phi_1(x) + \cdots + f \circ \phi_\ell(x) \\ &= s\, f(\gamma) \neq t\, f(\gamma) \\ &= f \circ \phi_1(y) + \cdots + f \circ \phi_\ell(y) = F(y). \end{aligned}$$

This concludes the proof. □

Proposition 6.2.1 follows directly from Proposition D.3.3.

*Proof of Proposition 6.2.1.* Note that, by Theorem 4.4.2, $\rho(\mathcal{U}_\sigma(C,I)) = \mathcal{C}_{S_n}(\mathbb{R}^n)$ 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 $C$ are $e_1^\top, \dots, e_n^\top$. Hence, by Proposition D.3.3, $(x,y) \in \rho(\mathcal{U}_\sigma(C,I))$ 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, $\mathcal{U}_\sigma(C,I)$ also has maximal separation power, and hence

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

Since

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

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 6.1.4.* By Proposition D.3.3, separation-constrained universality is equivalent to the ability to approximate any function of the form $F(\phi_1,\dots,\phi_\ell)$, where $F$ is continuous and $S_\ell$-invariant.

Recall that the basis maps are defined as

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

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 $i = 1,\dots,\ell$.

In particular, the function

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

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

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

that is,

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

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 $i = 1, \dots, \ell$. Therefore, $F$ cannot be approximated by $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(M^{1\times h}, N^{h\times 1})$.

This follows because the differential operator $D_{c_1} \cdots D_{c_\ell}$ reduces the degree of each monomial in $G$ by at most $\ell$. Thanks to the hypothesis $a_i + \ell < a_{i+1}$ for each $i = 1, \dots, \ell$, 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 $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(M_{h,h}\, N)$.

By direct computation, the coefficients of the monomials of multidegree $(a_1 - s_1, \dots, a_\ell - s_\ell)$ 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 $s_1, \dots, s_\ell \in \{0, \dots, \ell\}$, $s_1 + \cdots + s_\ell = \ell$ and $i_1, \dots, i_r$ 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 $\bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(M_{h,h}\, N)$. □

*Proof of Theorem 6.3.2.* Define $V$, $W$, and $\iota : V \to W$ as in Corollary D.3.5, which states that

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

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

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

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,

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

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. □

*Proof of Theorem 6.1.2.* Note that

$$\mathcal{M}(\phi_1,\dots,\phi_\ell) = \mathcal{M}(\psi_1,\dots,\psi_\ell)$$

for basis maps $\phi_1,\dots,\phi_\ell$ and $\psi_1,\dots,\psi_\ell$ such that

$$\mathrm{Span}\{\phi^1,\dots,\phi^m\} = \mathrm{Span}\{\psi^1,\dots,\psi^m\},$$

since

$$L(\phi_i) = L(\psi_i)$$

for each $i = 1,\dots,\ell$. In particular,

$$\mathcal{M}(\phi_1,\dots,\phi_\ell)^G = \mathcal{M}(\psi_1,\dots,\psi_\ell)^G.$$

Thus, it suffices to restrict to a specific basis $\phi^1,\dots,\phi^m$ chosen as follows.

Consider the decomposition

$$\mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^Y) \cong \mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^{Y_1}) \oplus \cdots \oplus \mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^{Y_r}),$$

and construct a basis of $\lambda(M)$ by choosing, for each $i = 1,\dots,r$, a basis of $\mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^{Y_i})$, and embedding it in $\mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^Y)$ via the canonical inclusion induced by the direct sum decomposition. In particular, there exists a partition

$$\mathcal{I}_1 \sqcup \cdots \sqcup \mathcal{I}_r = [m]$$

such that

$$\mathrm{Span}\{\phi^j\}_{j\in\mathcal{I}_i} = \mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^{Y_i})$$

for each $i = 1,\dots,r$, where each $\mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^{Y_i})$ is viewed as embedded in $\mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^Y)$. Equivalently, for each $j \in \mathcal{I}_i$ there exists $\psi^j \in \mathrm{Hom}_G(\mathbb{R}^X,\mathbb{R}^{Y_i})$ such that

$$\phi^j(x) = (0,\dots,0,\ \underbrace{\psi^j(x)}_{i\text{-th block}},0,\dots,0). \tag{D.7}$$

With this notation in place, we now prove that

$$\mathcal{M}(\phi_1,\dots,\phi_\ell)^G = \Delta(\phi_1,\dots,\phi_\ell).$$

Let $\mathcal{R} : \mathcal{C}(\mathbb{R}^n) \to \mathcal{C}(\mathbb{R}^n)^G$ be the Reynolds operator, namely

$$\mathcal{R}(F)(x) := \frac{1}{|G|}\sum_{g\in G} F(gx).$$

For each $F \in \mathcal{M}(\phi_1,\dots,\phi_\ell)$, write $F(x) = \sum_{k=1}^{\ell} f_k \circ \phi_k(x)$ for some continuous $f_1,\dots,f_\ell$. Moreover, since $\phi^1,\dots,\phi^m$ are $G$-equivariant, the induced $G$-action

permutes the corresponding family of coordinate maps. More precisely, for each $g \in G$ and each index $k$ we have

$$\phi_k(gx) = \phi_{g^{-1}\cdot k}(x) \qquad \text{for all } x.$$

Indeed, using equivariance of each $\phi^j$,

$$\begin{aligned}\big(\phi_k^1(gx), \dots, \phi_k^m(gx)\big) &= \big((g\phi^1)_k(x), \dots, (g\phi^m)_k(x)\big) \\ &= \big(\phi^1_{g^{-1}\cdot k}(x), \dots, \phi^m_{g^{-1}\cdot k}(x)\big),\end{aligned}$$

for each $k = 1, \dots, \ell$. Then

$$\begin{aligned}\mathcal{R}(F)(x) = \frac{1}{|G|}\sum_{g\in G} F(gx) &= \frac{1}{|G|}\sum_{g\in G}\sum_{k=1}^{\ell} f_k \circ \phi_k(gx). \\ &= \frac{1}{|G|}\sum_{g\in G}\sum_{k=1}^{\ell} f_k \circ \phi_{g\cdot k}(x).\end{aligned}$$

Next, group together indices belonging to the same block. For each $i = 1, \dots, r$, let $\pi_i : \mathbb{R}^m \to \mathbb{R}^{|\mathcal{I}_i|}$ be the coordinate projection induced by the partition $[m] = \mathcal{I}_1 \sqcup \cdots \sqcup \mathcal{I}_r$. That is,

$$\pi_i(z_1, \dots, z_m) := (z_j)_{j\in\mathcal{I}_i}.$$

In particular, for each $x$ we have

$$\pi_i\big(\phi_h^1(x), \dots, \phi_h^m(x)\big) = \big(\phi_h^j(x)\big)_{j\in\mathcal{I}_i} = \big(\psi_h^j(x)\big)_{j\in\mathcal{I}_i}.$$

Define

$$F_i := \frac{1}{|G|}\sum_{j\in\mathcal{I}_i} f_j \circ \pi_i, \qquad \text{and} \qquad H(y_1, \dots, y_r) := \sum_{i=1}^{r} F_i(y_i).$$

Then, using the block form in D.7,

$$\begin{aligned}\mathcal{R}(F) &= \frac{1}{|G|}\sum_{g\in G}\sum_{k=1}^{\ell} f_k \circ \phi_{g\cdot k} \\ &= \sum_{i=1}^{r}\sum_{j\in\mathcal{I}_i} F_i \circ \psi_j \\ &= \sum_{i=1}^{r}\sum_{j\in\mathcal{I}_i} H \circ \phi_j \\ &= \sum_{j=1}^{\ell} H \circ \phi_j.\end{aligned}$$

The right-hand side is a sum of terms of the form $\sum_{j=1}^{\ell} H \circ \phi_j$, hence it belongs to $\Delta(\phi_1, \dots, \phi_\ell)$. Moreover, the preceding regrouping shows that $\mathcal{R}(F)$ can be written in this form, so $\mathcal{R}(F) \in \Delta(\phi_1, \dots, \phi_\ell)$. Therefore,

$$\mathcal{M}(\phi_1, \dots, \phi_\ell)^G \subseteq \Delta(\phi_1, \dots, \phi_\ell).$$

Conversely, since $\Delta(\phi_1, \dots, \phi_\ell)$ is $G$-invariant by construction, we have

$$\Delta(\phi_1, \dots, \phi_\ell)^G = \Delta(\phi_1, \dots, \phi_\ell) \subseteq \mathcal{M}(\phi_1, \dots, \phi_\ell).$$

Combining the two inclusions yields

$$\mathcal{M}(\phi_1, \dots, \phi_\ell)^G = \Delta(\phi_1, \dots, \phi_\ell).$$

Finally, if $f$ is an invariant function, then by Theorem B.0.4 and Theorem D.3.2 we have

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

This concludes the proof. □

*Proof of Theorem 6.1.3.* The final part of the proof of Theorem 6.1.2 implies that if $f \in \mathcal{U}_\sigma(M, N)$, then for any $P \in \mathcal{I}(L(\phi_1) \cup \dots \cup L(\phi_\ell))$, $P(\partial_1, \dots, \partial_n) f = 0$. By Remark A.0.5, we know

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

For any $\alpha = 1, \dots, \ell$ 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. □

**Lemma D.3.4.** *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*

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

*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 $\iota^*$ is linear and preserves equivariance. Moreover, since $\iota$ is injective, the induced map $\iota^*$ 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 $\iota^*$ 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 D.3.5.** *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*

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

*for any $G$-representation $Z$. Then*

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

*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 $\iota$ and Lemma D.3.4, 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

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

since $\iota$ is precomposed with the input in the first layer. □

## D.4 Deep Equivariant Networks

### D.4.1 Universality in the Invariant Case

**Definition D.4.1.** Let $M_1, \dots, M_d$ be layer spaces. Let $\mathcal{B}_i$ be bases for the layer space $M_i$, and define

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

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 D.4.2.** *In the notation of Definition D.4.1,*

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

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

*Proof of Lemma D.4.2.* By the continuity of the parametrization map and the density of $M_i^{\mathbb{Q}}$ in $M_i$. □

Lemma D.4.2 implies the following corollary.

**Corollary D.4.3.** *Let $M_1, \dots, M_d$ be layer spaces with complete bias. There exists a countable family of $\mathcal{F} = \{f_h\}_{h \in \mathbb{N}} \subseteq \mathcal{N}_\sigma(M_1, \dots, M_d)$ such that*

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

**Lemma D.4.4.** *Let $V = \mathbb{R}^d$ with its usual topology and let $\rho$ 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 D.4.4.* 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 $\phi$ is continuous and bijective, and since $K/\rho$ is compact Hausdorff and $\widehat{F}(K)$ is Hausdorff, $\phi$ 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

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

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. □

### D.4.2 From Invariance to Equivariance

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 $\pi_x$, 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 D.4.5.** *If the action of $G$ on $X$ is transitive, then reconstruction map $\theta_x^*$ is a well-defined, continuous linear operator such that*

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

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

*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

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

where in the second equality we use the fact that $g^{-1}g_1, \ldots, g^{-1}g_t$ is another transversal for $G/G_x$. These observations prove that $\theta_x^*$ 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 $g_1, \ldots, g_t$ is a transversal of $G/G_x$ if and only if $g_1 x, \ldots, g_t x$ is the $G$-orbit of $x$. Thus for each $f \in \mathcal{C}(V, \mathbb{R}^X)$ we can write

$$f(v) = \sum_{i=1}^{t} \pi_{g_i x} f(v) e_{g_i x}. \tag{D.8}$$

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

$$\begin{aligned}\theta_x^* \pi_{x^*} f(v) &= \sum_{i=1}^{t} \pi_x f(g_i^{-1}v)\, e_{g_i x} \\ &= \sum_{i=1}^{t} \pi_x\big(g_i^{-1} \cdot f(v)\big)\, e_{g_i x} \\ &= \sum_{i=1}^{t} \pi_{g_i x} f(v)\, e_{g_i x} \\ &\overset{\text{Equation D.8}}{=} f(v).\end{aligned}$$

□

Proposition D.4.5 says that $\pi_{x^*}$ 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 4.4.

*Proof of Proposition 5.1.2.* The claim follows directly from Proposition D.4.5: applying $\pi_x$ yields one inclusion, while the reconstruction map yields the other. □

*Remark* D.4.6 (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 extend, with minor modifications, to the non-transitive case, which we address next. Let $X = Y_1 \sqcup \cdots \sqcup Y_s$ be the decomposition of $X$ into $G$-orbits. For each $i=1,\dots,s$, denote by $\pi_{\mathbb{R}^{Y_i}}:\mathbb{R}^X \to \mathbb{R}^{Y_i}$ the standard projection. Consider the maps

$$\Phi: \begin{array}{c} \mathcal{C}_G(V,\mathbb{R}^X) \to \mathcal{C}_G(V,\mathbb{R}^{Y_1}) \oplus \cdots \oplus \mathcal{C}_G(V,\mathbb{R}^{Y_s}) \\ f \mapsto (\pi_{\mathbb{R}^{Y_1}} f, \dots, \pi_{\mathbb{R}^{Y_s}} f), \end{array}$$

and

$$\Psi: \begin{array}{c} \mathcal{C}_G(V,\mathbb{R}^{Y_1}) \oplus \cdots \oplus \mathcal{C}_G(V,\mathbb{R}^{Y_s}) \to \mathcal{C}_G(V,\mathbb{R}^X) \\ (f_1,\dots,f_s) \mapsto \big[x \mapsto (f_1(x),\dots,f_s(x))\big]. \end{array}$$

Note that $\Phi$ is a homeomorphism onto its image. Let $x_1,\dots,x_s$ be points with $x_i \in Y_i$. Then, for each $i=1,\dots,s$, we have

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

Moreover, for each $i=1,\dots,s$,

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

Therefore, understanding $\mathcal{C}_G(V,\mathbb{R}^X)$ reduces to understanding $\mathcal{C}_{G_{x_i}}(V,\mathbb{R})$ for each $i=1,\dots,s$, and then reconstructing $\mathcal{C}_G(V,\mathbb{R}^X)$ via $\theta^*_{x_1},\dots,\theta^*_{x_s}$ and $\Psi$. In particular, since we focus on closed linear subspaces $\mathcal{U} \subseteq \mathcal{C}_G(V,\mathbb{R}^X)$, it is enough to study the subspaces $\pi_{x_i^*}\mathcal{U}$ for $i=1,\dots,s$, namely universality classes of invariant neural networks.

**Proposition D.4.7.** *The following equality is true:*

$$\mathcal{U}_\sigma(\underbrace{C,\dots,C}_{d\ times}) = \{(x_1,\dots,x_n)\mapsto(f(x_1),\dots,f(x_n)) \mid f\in\mathcal{C}(\mathbb{R})\} \subsetneq \mathcal{C}_{S_n}(\mathbb{R}^n,\mathbb{R}^n). \tag{D.9}$$

*Proof of Proposition D.4.7.* 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 $\lambda(C) = \mathrm{Span}\,\{x\mapsto id_{\mathbb{R}^X}\cdot x\}$. Elements in $C^{1,h}$ can be represented as affine maps $x\mapsto Bx+c$ where $B$ and $c$ have the following block representations

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

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

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

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

$$\theta_{s,i}(x) = \sum_{j=1}^{h} a_{s,j}\sigma\left(b_j x_i + c_j\right) = \xi_s(x) \tag{D.10}$$

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. □

### D.4.3 Universality in the Equivariant Case

**Lemma D.4.8.** *Let $X, Y, Z$ be metric spaces, and let $\mathcal{G} \subseteq \mathcal{C}(X, Y)$ and $\mathcal{F} \subseteq \mathcal{C}(Y, Z)$. The following identities hold.*

$$\rho(\overline{\mathcal{G}} \mathbin{\hat{\circ}} \mathcal{F}) = \rho(\mathcal{G} \mathbin{\hat{\circ}} \mathcal{F}) = \rho(\mathcal{G} \mathbin{\hat{\circ}} \overline{\mathcal{F}}),$$

*where the closure is taken with respect to uniform convergence on compact sets.*

*Proof.* For convenience, recall the definition

$$\mathcal{G} \mathbin{\hat{\circ}} \mathcal{F} := \{f \circ g \mid f \in \mathcal{F},\, g \in \mathcal{G}\}.$$

We start by proving the first equality. Since $\mathcal{G} \mathbin{\hat{\circ}} \mathcal{F} \subseteq \overline{\mathcal{G}} \mathbin{\hat{\circ}} \mathcal{F}$, we have

$$\rho(\overline{\mathcal{G}} \mathbin{\hat{\circ}} \mathcal{F}) \subseteq \rho(\mathcal{G} \mathbin{\hat{\circ}} \mathcal{F}).$$

For the reverse inclusion, let $(x, y) \in \rho(\mathcal{G} \mathbin{\hat{\circ}} \mathcal{F})$, that is,

$$f \circ g(x) = f \circ g(y) \qquad \text{for each } f \in \mathcal{F} \text{ and } g \in \mathcal{G}.$$

Fix $f \in \mathcal{F}$ and let $g \in \overline{\mathcal{G}}$. By definition of the closure, there exists a sequence $(g_n)_{n\in\mathbb{N}} \subseteq \mathcal{G}$ such that $g_n \to g$ uniformly on compact sets. For each $n \in \mathbb{N}$,

$$f(g_n(x)) = f(g_n(y)),$$

and by continuity of $f$ we obtain $f(g(x)) = f(g(y))$. Since $f \in \mathcal{F}$ and $g \in \overline{\mathcal{G}}$ were arbitrary,

$$(x, y) \in \rho(\overline{\mathcal{G}} \,\hat{\circ}\, \mathcal{F}),$$

i.e.,

$$\rho(\mathcal{G} \,\hat{\circ}\, \mathcal{F}) \subseteq \rho(\overline{\mathcal{G}} \,\hat{\circ}\, \mathcal{F}).$$

This proves the first equality.

We now prove the second equality. Since $\mathcal{G} \,\hat{\circ}\, \mathcal{F} \subseteq \mathcal{G} \,\hat{\circ}\, \overline{\mathcal{F}}$, we have

$$\rho(\mathcal{G} \,\hat{\circ}\, \overline{\mathcal{F}}) \subseteq \rho(\mathcal{G} \,\hat{\circ}\, \mathcal{F}).$$

For the reverse inclusion, let $(x, y) \in \rho(\mathcal{G} \,\hat{\circ}\, \mathcal{F})$, that is,

$$f \circ g(x) = f \circ g(y) \qquad \text{for each } f \in \mathcal{F} \text{ and } g \in \mathcal{G}.$$

Fix $g \in \mathcal{G}$ and let $f \in \overline{\mathcal{F}}$. By definition of the closure, there exists a sequence $(f_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}$ such that $f_n \to f$ uniformly on compact sets. For each $n \in \mathbb{N}$,

$$f_n(g(x)) = f_n(g(y)),$$

hence taking limits yields $f(g(x)) = f(g(y))$. Since $g \in \mathcal{G}$ and $f \in \overline{\mathcal{F}}$ were arbitrary,

$$(x, y) \in \rho(\mathcal{G} \,\hat{\circ}\, \overline{\mathcal{F}}),$$

i.e.,

$$\rho(\mathcal{G} \,\hat{\circ}\, \mathcal{F}) \subseteq \rho(\mathcal{G} \,\hat{\circ}\, \overline{\mathcal{F}}).$$

This concludes the proof. □

**Lemma D.4.9.** *Let $M$ be a layer space. Then for each $k, h \in \mathbb{N}_0$ we have*

$$C^{k\times 1} \,\hat{\circ}\, P^{1\times h} \subseteq P^{k\times h}.$$

*Proof.* Let $\phi \in C^{k\times 1}$. Then $\phi$ can be written as

$$\phi(x_1, \ldots, x_k) = \phi_1(x_1) + \cdots + \phi_k(x_k)$$

for some $\phi_1, \ldots, \phi_k \in C$.
Let $\psi \in P^{1\times h}$. Then $\psi$ can be written as

$$\psi(x) = \big(\psi_1(x), \ldots, \psi_h(x)\big)$$

for some $\psi_1, \ldots, \psi_h \in P$.
The composition $\psi \circ \phi \in C^{k\times 1} \,\hat{\circ}\, M^{1\times h}$ can be written

$$\psi \circ \phi(x_1, \ldots, x_k) = \big(\sum_{j=1}^{k} \psi_i \circ \phi_j(x_j)\big)_{i,j}$$

where $\psi_i \circ \phi_j \in M$ for each $i$ and $j$ since each $\phi_j$ is a scalar multiple of the identity plus a translational part. □

**Lemma D.4.10.** *The projection $\pi_i^* \mathcal{U}_\sigma(C,P,C)$ separates $\mathrm{Stab}_{S_n}(i)$-orbits in $\mathbb{R}^n$.*

*Proof.* Without loss of generality, assume $i=1$, and note that

$$\pi_1^* \mathcal{U}_\sigma(C,P,C) = \mathcal{U}_\sigma(C,P,P_1),$$

where $P_1 := \pi_1^* C$. By Lemma D.4.8, we obtain

$$\bigcup_{k\in\mathbb{N}} \mathcal{N}_\sigma(C^{1\times k}, C^{k\times 1}) \mathbin{\hat{\circ}} \bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(P^{1\times h}, P_1^{h\times 1}) \subseteq \bigcup_{k,h\in\mathbb{N}} \mathcal{N}_\sigma(C^{1\times k}, P^{k\times h}, P_1^{h\times 1}).$$

Then, by Lemma D.4.9, we have

$$\begin{aligned}
\rho\big(\mathcal{U}_\sigma(C,P,P_1)\big) &= \rho\Big(\bigcup_{k,h\in\mathbb{N}} \mathcal{N}_\sigma(C^{1\times k}, P^{k\times h}, P_1^{h\times 1})\Big) \\
&\subseteq \rho\Big(\bigcup_{k\in\mathbb{N}} \mathcal{N}_\sigma(C^{1\times k}, C^{k\times 1}) \mathbin{\hat{\circ}} \bigcup_{h\in\mathbb{N}} \mathcal{N}_\sigma(P^{1\times h}, P_1^{h\times 1})\Big) \\
&= \rho\big(\mathcal{U}_\sigma(C,C) \mathbin{\hat{\circ}} \mathcal{U}_\sigma(P,P_1)\big).
\end{aligned}$$

Since functions in $\mathcal{U}_\sigma(C,P,P_1)$ are $\mathrm{Stab}_{S_n}(1)$-invariant, they can at most separate $\mathrm{Stab}_{S_n}(1)$-orbits. Therefore, it suffices to show that $\mathcal{U}_\sigma(C,C) \mathbin{\hat{\circ}} \mathcal{U}_\sigma(P,P_1)$ separates these orbits. We know by Proposition D.4.7 that

$$\mathcal{U}_\sigma(C,C) = \Big\{ (x_1,\dots,x_n) \mapsto (f(x_1),\dots,f(x_n)) \;\Big|\; f \in \mathcal{C}(\mathbb{R}) \Big\}.$$

We now characterize elements in $\mathcal{U}_\sigma(P,P_1)$. Before proceeding, recall that

$$C = \{\, x \mapsto \lambda x + \mu \mathbb{1} \mid \lambda,\mu \in \mathbb{R} \,\}.$$

Then

$$P_1 = \pi_1^* C = \{\, x \mapsto \pi_1(\lambda x + \mu\mathbb{1}) \mid \lambda,\mu\in\mathbb{R}\,\} = \{\, x \mapsto \lambda e_1^\top x + \mu \mid \lambda,\mu\in\mathbb{R}\,\}.$$

Elements in $\mathcal{U}_\sigma(P,P_1)$ are limits of functions of the form

$$\eta(x) = A \circ \tilde{\sigma} \circ B(x),$$

where

$$B = \begin{bmatrix} b_{1,1}\ \mathrm{id} + b_{1,2}\ \mathbb{1}^\top\mathbb{1} \\ \vdots \\ b_{h,1}\ \mathrm{id} + b_{h,2}\ \mathbb{1}^\top\mathbb{1} \end{bmatrix}, \qquad A = \begin{bmatrix} a_1 \cdot e_1^\top & \cdots & a_h \cdot e_h^\top \end{bmatrix},$$

for arbitrary $h \geq 1$. Then, for $x = (x_1,\dots,x_n)$,

$$\eta(x) = \sum_{r=1}^{h} a_r\, \sigma\Big(b_{r,1}x_1 + b_{r,2}(x_1+\cdots+x_n)\Big) = \zeta\big(x_1, x_1+\cdots+x_n\big),$$

where $\zeta \in \mathcal{N}_\sigma(L^{2\times h}, L^{h\times 1})$. Hence,

$$\mathcal{U}_\sigma(P, P_1) = \Big\{(x_1, \dots, x_n) \mapsto f\big(x_1, x_1 + \cdots + x_n\big) \;\Big|\; f \in \mathcal{C}(\mathbb{R}^2)\Big\}.$$

Therefore,

$$\mathcal{U} \coloneqq \mathcal{U}_\sigma(C, C) \,\hat{\circ}\, \mathcal{U}_\sigma(P, P_1) = \Big\{(x_1, \dots, x_n) \mapsto f\big(g(x_1),\, g(x_1) + \cdots + g(x_n)\big) \;\Big|\; f \in \mathcal{C}(\mathbb{R}^2),\ g \in \mathcal{C}(\mathbb{R})\Big\}.$$

Note that the family

$$\Big\{(x_1, \dots, x_n) \mapsto g(x_1) + \cdots + g(x_n) \;\Big|\; g \in \mathcal{C}(\mathbb{R})\Big\}$$

separates $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ if and only if there exists a permutation $\gamma \in S_n$ such that $x = \gamma y$. Moreover, $\mathcal{U}$ separates $x$ and $y$ whenever $x_1 \neq y_1$, for instance by choosing $g = \mathrm{id}$ and $f(u, v) = u$. Thus, $\mathcal{U}$ separates $x$ and $y$ if and only if there exists a permutation $\gamma \in \mathrm{Stab}_{S_n}(1)$ such that $x = \gamma y$. This concludes the proof. □

# Bibliography


[1] Y. Alexander, Y. Slutzky, Y. Ran-Milo, and N. Cohen. Do Neural Networks Need Gradient Descent to Generalize? A Theoretical Study. Oct. 2025.

[2] E. Alsentzer, S. G. Finlayson, M. M. Li, and M. Zitnik. Subgraph Neural Networks, Nov. 2020. arXiv:2006.10538 [cs, stat].

[3] A. Andreyev and P. Beneventano. Edge of Stochastic Stability: Revisiting the Edge of Stability for SGD, Dec. 2025. arXiv:2412.20553 [cs].

[4] M. F. Atiyah and I. G. MacDonald. *Introduction To Commutative Algebra*. Avalon Publishing, Feb. 1994. Google-Books-ID: HOASFid4x18C.

[5] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [cs, stat]*, Oct. 2018. arXiv: 1806.01261.

[6] C. Battiloro, E. Karaismailoğlu, M. Tec, G. Dasoulas, M. Audirac, and F. Dominici. E(n) Equivariant Topological Neural Networks, Feb. 2025. arXiv:2405.15429 [cs].

[7] S. Beddar-Wiesing, G. A. D'Inverno, C. Graziani, V. Lachi, A. Moallemy-Oureh, F. Scarselli, and J. M. Thomas. Weisfeiler–Lehman goes dynamic: An analysis of the expressive power of Graph Neural Networks for attributed and dynamic graphs. *Neural Networks*, 173:106213, May 2024.

[8] A. Behboodi, G. Cesa, and T. S. Cohen. A PAC-Bayesian Generalization Bound for Equivariant Networks. *Advances in Neural Information Processing Systems*, 35:5654–5668, Dec. 2022.

[9] E. J. Bekkers, S. Vadgama, R. D. Hesselink, P. A. van der Linden, and D. W. Romero. Fast, Expressive SE$(n)$ Equivariant Networks through Weight-Sharing in Position-Orientation Space, Oct. 2023. arXiv:2310.02970 [cs, math].

[10] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *arXiv:1812.11118 [cs, stat]*, Sept. 2019. arXiv: 1812.11118.

[11] P. Beneventano. On the Trajectories of SGD Without Replacement, Apr. 2024. arXiv:2312.16143 [cs].

[12] G. Benkart, T. Halverson, and N. Harman. Dimensions of irreducible modules for partition algebras and tensor power multiplicities for symmetric and alternating groups, May 2016. arXiv:1605.06543 [math].

[13] E. Berman, J. Ginesin, M. Pacini, and R. Walters. On Uncertainty Calibration for Equivariant Functions. *Transactions on Machine Learning Research*, Oct. 2025.

[14] B. Bevilacqua, F. Frasca, D. Lim, B. Srinivasan, C. Cai, G. Balamurugan, M. M. Bronstein, and H. Maron. Equivariant Subgraph Aggregation Networks, Mar. 2022. arXiv:2110.02910 [cs, stat].

[15] F. M. Bianchi and V. Lachi. The expressive power of pooling in Graph Neural Networks. *Advances in Neural Information Processing Systems*, 36:71603–71618, Dec. 2023.

[16] B. Blum-Smith and S. Villar. Machine learning and invariant theory. *Notices of the American Mathematical Society*, 70(08):1, Sept. 2023. arXiv:2209.14991 [stat].

[17] A. Bogatskiy, B. Anderson, J. Offermann, M. Roussi, D. Miller, and R. Kondor. Lorentz Group Equivariant Neural Network for Particle Physics. In *Proceedings of the 37th International Conference on Machine Learning*, pages 992–1002. PMLR, Nov. 2020.

[18] S. Bortolotti, E. Marconato, T. Carraro, P. Morettin, E. v. Krieken, A. Vergari, S. Teso, and A. Passerini. A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts. Nov. 2024.

[19] L. Bottou and O. Bousquet. The Tradeoffs of Large Scale Learning. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[20] L. Braun, C. Dominé, J. Fitzgerald, and A. Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, 35:6615–6629, Dec. 2022.

[21] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [cs, stat]*, May 2021. arXiv: 2104.13478.

[22] O. Calin. *Deep Learning Architectures: A Mathematical Approach*. Springer Publishing Company, Incorporated, 1st edition, 2020.

[23] P.-y. Chiang, R. Ni, D. Y. Miller, A. Bansal, J. Geiping, M. Goldblum, and T. Goldstein. Loss Landscapes are All You Need: Neural Network Generalization Can Be Explained Without the Implicit Bias of Gradient Descent. Sept. 2022.

[24] T. Cohen. *Equivariant convolutional networks*. PhD Thesis, Taco Cohen, 2021.

[25] T. Cohen and M. Welling. Group Equivariant Convolutional Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2990–2999. PMLR, June 2016.

[26] T. S. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling. Gauge Equivariant Convolutional Networks and the Icosahedral CNN, May 2019. arXiv:1902.04615 [cs, stat].

[27] T. S. Cohen and M. Welling. Steerable CNNs. In *International Conference on Learning Representations*, Nov. 2016.

[28] N. Dym and S. J. Gortler. Low Dimensional Invariant Embeddings for Universal Geometric Learning, May 2022. arXiv:2205.02956 [cs, math].

[29] N. Dym, H. Lawrence, and J. W. Siegel. Equivariant Frames and the Impossibility of Continuous Canonicalization, June 2024. arXiv:2402.16077 [cs].

[30] A. S. d'Avila Garcez, K. B. Broda, and D. M. Gabbay. *Neural-Symbolic Learning Systems*. Perspectives in Neural Computing. Springer, London, 2002.

[31] W. E, J. Han, and A. Jentzen. Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations. *Communications in Mathematics and Statistics*, 5(4):349–380, Dec. 2017.

[32] B. Elesedy. Group Symmetry in PAC Learning. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

[33] F. Ferrini, V. Lachi, A. Longa, B. Lepri, and A. Passerini. GNNs Meet Sequence Models Along the Shortest-Path: an Expressive Method for Link Prediction. Oct. 2025.

[34] B. Finkelshtein, C. Baskin, H. Maron, and N. Dym. A Simple and Universal Rotation Equivariant Point-Cloud Network. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, pages 107–115. PMLR, Nov. 2022.

[35] M. Finzi, G. Benton, and A. G. Wilson. Residual Pathway Priors for Soft Equivariance Constraints. In *Advances in Neural Information Processing Systems*, volume 34, pages 30037–30049. Curran Associates, Inc., 2021.

[36] P. Flor. On groups of non-negative matrices. *Compositio Mathematica*, 21(4):376–382, 1969.

[37] F. Frasca, B. Bevilacqua, M. M. Bronstein, and H. Maron. Understanding and Extending Subgraph GNNs by Rethinking Their Symmetries, Oct. 2022. arXiv:2206.11140 [cs].

[38] F. Fuchs, D. Worrall, V. Fischer, and M. Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

[39] W. Fulton and J. Harris. *Representation Theory*, volume 129 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2004.

[40] J. Gasteiger, F. Becker, and S. Günnemann. GemNet: Universal Directional Graph Neural Networks for Molecules, Apr. 2022. arXiv:2106.08903 [physics, stat].

[41] F. Geerts. The expressive power of kth-order invariant graph networks, July 2020. arXiv:2007.12035 [cs, math, stat].

[42] F. Geerts, T. Muñoz, C. Riveros, and D. Vrgoč. Expressive Power of Linear Algebra Query Languages. In *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS'21, pages 342–354, New York, NY, USA, 2021. Association for Computing Machinery.

[43] F. Geerts and J. L. Reutter. Expressiveness and Approximation Properties of Graph Neural Networks. *Preprint*, page 43, 2022.

[44] M. Geiger and T. Smidt. e3nn: Euclidean Neural Networks, July 2022. arXiv:2207.09453 [cs].

[45] G. Gidel, F. Bach, and S. Lacoste-Julien. Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[46] 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.

[47] Grubb, G. *Distributions and Operators*, volume 252 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2009.

[48] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit Bias of Gradient Descent on Linear Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[49] B. Hu, D. Wang, D. Klee, H. Tian, X. Zhu, H. Huang, R. Platt, and R. Walters. 3D Equivariant Visuomotor Policy Learning via Spherical Projection. Oct. 2025.

[50] H. Huang, H. Liu, D. Wang, R. Walters, and R. Platt. Match Policy: A Simple Pipeline from Point Cloud Registration to Manipulation Policies. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16907–16914, May 2025.

[51] A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[52] Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1772–1798. PMLR, June 2019.

[53] M. Jia, H. Huang, Z. Zhang, C. Wang, L. Zhao, D. Wang, J. X. Liu, R. Walters, R. Platt, and S. Tellex. Learning Efficient and Robust Language-Conditioned Manipulation Using Textual-Visual Relevancy and Equivariant Language Mapping. *IEEE Robotics and Automation Letters*, 10(8):8204–8211, Aug. 2025.

[54] C. K. Joshi, C. Bodnar, S. V. Mathis, T. Cohen, and P. Lio. On the Expressive Power of Geometric Graph Neural Networks. *International Conference of Learning Representations*, 2023.

[55] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, Aug. 2021.

[56] S.-O. Kaba, A. K. Mondal, Y. Zhang, Y. Bengio, and S. Ravanbakhsh. Equivariance with Learned Canonicalization Functions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15546–15566. PMLR, July 2023.

[57] R. Kakarala. *Triple correlation on groups*. phd, University of California at Irvine, USA, 1992. UMI Order No. GAX93-04094.

[58] P. Kannappan. Functional Equations from Information Theory. In P. Kannappan, editor, *Functional Equations and Inequalities with Applications*, Springer Monographs in Mathematics, pages 403–467. Springer US, Boston, MA, 2009.

[59] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, June 2021.

[60] B. T. Kiani, J. Wang, and M. Weber. Hardness of Learning Neural Networks under the Manifold Hypothesis. *Advances in Neural Information Processing Systems*, 37:5661–5696, Dec. 2024.

[61] T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, Feb. 2017. arXiv: 1609.02907.

[62] G. Kiss and M. Laczkovich. Linear functional equations. 2014.

[63] R. Kondor. N-body Networks: a Covariant Hierarchical Neural Network Architecture for Learning Atomic Potentials, Mar. 2018. arXiv:1803.01588 [cs].

[64] R. Kondor and S. 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.

[65] V. Lachi, F. Ferrini, A. Longa, B. Lepri, A. Passerini, and M. Jaeger. Bridging Theory and Practice in Link Representation with Graph Neural Networks. Oct. 2025.

[66] M. W. Lafarge, E. J. Bekkers, J. P. W. Pluim, R. Duits, and M. Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis*, 68:101849, Feb. 2021.

[67] I. Lagaris, A. Likas, and D. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, Sept. 1998.

[68] L. Lang and M. Weiler. A Wigner-Eckart Theorem for Group Equivariant Convolution Kernels. In *International Conference on Learning Representations*, 2021.

[69] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. Number: 7553.

[70] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998. Conference Name: Proceedings of the IEEE.

[71] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[72] J. Ling, A. Kurzawski, and J. Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, Nov. 2016.

[73] L. Lovász. Large Networks and Graph Limits. volume 60 of *Colloquium Publications*, Providence, Rhode Island, Dec. 2012. American Mathematical Society.

[74] W. J. Maddox, G. Benton, and A. G. Wilson. Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited, May 2020. arXiv:2003.02139 [cs].

[75] E. Marconato, S. Bortolotti, E. v. Krieken, A. Vergari, A. Passerini, and S. Teso. BEARS Make Neuro-Symbolic Models Aware of their Reasoning Shortcuts. June 2024.

[76] E. Marconato, S. Lachapelle, S. Weichwald, and L. Gresele. All or None: Identifiable Linear Properties of Next-Token Predictors in Language Modeling. Feb. 2025.

[77] H. Maron, H. Ben-Hamu, H. Serviansky, and Y. Lipman. Provably Powerful Graph Networks. *International Conference of Learning Representations*, 2019.

[78] H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman. Invariant and Equivariant Graph Networks. In *International Conference on Learning Representations*, Sept. 2018.

[79] H. Maron, E. Fetaya, N. Segol, and Y. Lipman. On the Universality of Invariant Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4363–4371. PMLR, May 2019.

[80] H. Maron, O. Litany, G. Chechik, and E. Fetaya. On Learning Sets of Symmetric Elements. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6734–6744. PMLR, Nov. 2020.

[81] P. Misof, P. Kessel, and J. E. Gerken. Equivariant Neural Tangent Kernels. June 2025.

[82] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4602–4609, July 2019.

[83] M. Pacini, X. Dong, B. Lepri, and G. Santin. A Characterization Theorem for Equivariant Networks with Point-wise Activations. In *The Twelfth International Conference on Learning Representations*, Oct. 2023.

[84] M. Pacini, X. Dong, B. Lepri, and G. Santin. Separation Power of Equivariant Neural Networks. In *The Thirteenth International Conference on Learning Representations*, Oct. 2024.

[85] M. Pacini, M. Petrache, B. Lepri, S. Trivedi, and R. Walters. On Universality of Deep Equivariant Networks, Oct. 2025. arXiv:2510.15814 [stat].

[86] M. Pacini, G. Santin, B. Lepri, and S. Trivedi. On Universality Classes of Equivariant Networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, Oct. 2025.

[87] A. Peleg and M. Hein. Bias of Stochastic Gradient Descent or the Architecture: Disentangling the Effects of Overparameterization of Neural Networks. June 2024.

[88] S. Pesme, L. Pillaud-Vivien, and N. Flammarion. Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity. Nov. 2021.

[89] M. Petrache and S. Trivedi. Approximation-Generalization Trade-offs under (Approximate) Group Equivariance. *Advances in Neural Information Processing Systems*, 36:61936–61959, Dec. 2023.

[90] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, Jan. 1999.

[91] A. Pinkus. *Ridge Functions*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2015.

[92] G. P. P. Pun, R. Batra, R. Ramprasad, and Y. Mishin. Physically informed artificial neural networks for atomistic modeling of materials. *Nature Communications*, 10(1):2339, May 2019.

[93] O. Puny, M. Atzmon, E. J. Smith, I. Misra, A. Grover, H. Ben-Hamu, and Y. Lipman. Frame Averaging for Invariant and Equivariant Network Design. Oct. 2021.

[94] O. Puny, D. Lim, B. T. Kiani, H. Maron, and Y. Lipman. Equivariant Polynomials for Graph Neural Networks, June 2023. arXiv:2302.11556 [cs].

[95] C. 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.

[96] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Inferring solutions of differential equations using noisy multi-fidelity data. *Journal of Computational Physics*, 335:736–746, Apr. 2017.

[97] M. Rama, G. Santin, G. Cencetti, M. Tizzoni, and B. Lepri. Forecasting Seasonal Influenza Epidemics with Physics-Informed Neural Networks, June 2025. arXiv:2506.03897 [physics].

[98] S. Ravanbakhsh. Universal Equivariant Multilayer Perceptrons. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7996–8006. PMLR, Nov. 2020.

[99] S. Ravanbakhsh, J. Schneider, and B. Póczos. Equivariance Through Parameter-Sharing. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2892–2901. PMLR, July 2017.

[100] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic Routing Between Capsules. *Advances in Neural Information Processing Systems*, 2017.

[101] B. E. Sagan. *The Symmetric Group*, volume 203 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 2001.

[102] A. Sannai, M. Imaizumi, and M. Kawano. Improved generalization bounds of group invariant / equivariant deep networks via quotient feature spaces. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 771–780. PMLR, Dec. 2021.

[103] A. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Dec. 2013.

[104] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini. Computational Capabilities of Graph Neural Networks. *IEEE Transactions on Neural Networks*, 20(1):81–102, Jan. 2009.

[105] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *Faculty of Informatics - Papers (Archive)*, Jan. 2009.

[106] N. Segol and Y. Lipman. On Universal Equivariant Set Networks, Jan. 2020. arXiv:1910.02421 [cs, stat].

[107] M. Seleznova and G. Kutyniok. Analyzing Finite Neural Networks: Can We Trust Neural Tangent Kernel Theory? In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pages 868–895. PMLR, Apr. 2022.

[108] M. Seleznova and G. Kutyniok. Neural Tangent Kernel Beyond the Infinite-Width Limit: Effects of Depth and Initialization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19522–19560. PMLR, June 2022.

[109] M. Seleznova, D. Weitzner, R. Giryes, G. Kutyniok, and H.-H. Chou. Neural (Tangent Kernel) Collapse. *Advances in Neural Information Processing Systems*, 36:16240–16270, Dec. 2023.

[110] L. Serafini, I. Donadello, and A. d. Garcez. Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation. In *Proceedings of the Symposium on Applied Computing*, SAC '17, pages

125–130, New York, NY, USA, Apr. 2017. Association for Computing Machinery.

[111] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, 2014.

[112] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735, May 1997. Conference Name: IEEE Transactions on Neural Networks.

[113] B. Tahmasebi and S. Jegelka. Generalization Bounds for Canonicalization: A Comparative Study with Group Averaging. Oct. 2024.

[114] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, May 2018. arXiv:1802.08219 [cs].

[115] Victor Garcia Satorras, E. Hoogeboom, and M. Welling. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9323–9332. PMLR, July 2021.

[116] S. Villar, D. W. Hogg, K. Storey-Fisher, W. Yao, and B. Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. In *Advances in Neural Information Processing Systems*, volume 34, pages 28848–28863. Curran Associates, Inc., 2021.

[117] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt. Equivariant Diffusion Policy. Sept. 2024.

[118] M. Weiler and G. Cesa. General E(2) - Equivariant Steerable CNNs. *Advances in Neural Information Processing Systems*, 2019.

[119] M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. 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.

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

[121] J. Wood and J. Shawe-Taylor. Representation theory and invariant neural networks. *Discrete Applied Mathematics*, 69(1-2):33–60, Aug. 1996.

[122] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic Networks: Deep Translation and Rotation Equivariance. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7168–7177, July 2017.

[123] Y. Xie and T. Smidt. A Tale of Two Symmetries: Exploring the Loss Landscape of Equivariant Models. Oct. 2025.

[124] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[125] B. Zhang, J. Gai, Y. Du, Q. Ye, D. He, and L. Wang. Beyond Weisfeiler-Lehman: A Quantitative Framework for GNN Expressiveness, Jan. 2024. arXiv:2401.08514 [cs, math].

[126] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. Feb. 2017.

[127] B. Zhao, N. Dehmamy, R. Walters, and R. Yu. Understanding Mode Connectivity via Parameter Space Symmetry. June 2025.

[128] B. Zhao, I. Ganev, R. Walters, R. Yu, and N. Dehmamy. Symmetries, Flat Minima, and the Conserved Quantities of Gradient Flow. Sept. 2022.

[129] A. Zweig and J. Bruna. A Functional Perspective on Learning Symmetric Functions with Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 13023–13032. PMLR, July 2021.