Universal Equivariant Multilayer Perceptrons

Siamak Ravanbakhsh [1 2]

Abstract

Group invariant and equivariant Multilayer Perceptrons (MLP), also known as Equivariant Networks and Group Group Convolutional Neural Networks (G-CNN) have achieved remarkable success in learning on a variety of data structures, such as sequences, images, sets, and graphs. This paper proves the universality of a broad class of equivariant MLPs with a single hidden layer. In particular, it is shown that having a hidden layer on which the group acts regularly is sufficient for universal equivariance (invariance). For example, some types of steerable-CNNs become universal. Another corollary is the unconditional universality of equivariant MLPs for all Abelian groups. A third corollary is the universality of equivariant MLPs with a high-order hidden layer, where we give both group-agnostic bounds and group-specific bounds on the order of the hidden layer that guarantees universal equivariance.

1. Introduction

Invariance and equivariance properties constrain the output of a function under various transformations of its input. This constraint serves as a strong learning bias that has proven useful in sample efficient learning for a wide range of structured data. In this work, we are interested in universality results for Multilayer Perceptrons (MLPs) that are constrained to be equivariant or invariant. This type of result guarantees that the model can approximate any continuous equivariant (invariant) function with an arbitrary precision, in the same way an unconstrained MLP can approximate an arbitrary continuous function (Hornik et al., 1989; Cybenko, 1989; Funahashi, 1989).

Study of invariance in neural networks goes back to the book of Perceptrons (Minsky & Papert, 2017), where the necessity of parameter-sharing for invariance was used to prove the limitation of a single layer Perceptron. The follow-up work showed how parameter symmetries can be used to achieve invariance to finite and infinite groups (Shawe-Taylor, 1989; Wood & Shawe-Taylor, 1996; Shawe-Taylor, 1993; Wood, 1996). These fundamental early works went unnoticed during the resurgence of neural network research and renewed attention to symmetry (Hinton et al., 2011; Mallat, 2012; Bruna & Mallat, 2013; Gens & Domingos, 2014; Jaderberg et al., 2015; Dieleman et al., 2016; Cohen & Welling, 2016a).

When equivariance constraints are imposed on feed-forward layers in an MLP, the linear maps in each layer is constrained to use tied parameters (Wood & Shawe-Taylor, 1996; Ravanbakhsh et al., 2017b). This model that we call an *equivariant MLP* appears in deep learning with sets (Zaheer et al., 2017; Qi et al., 2017), exchangeable tensors (Hartford et al., 2018), graphs (Maron et al., 2018), relational data (Graham & Ravanbakhsh, 2019), and sets of symmetric elements (Maron et al., 2020). Universality results for some of these models exists (Zaheer et al., 2017; Segol & Lipman, 2019; Keriven & Peyré, 2019; Maron et al., 2020). Broader results for high order *invariant* MLPs appears in (Maron et al., 2019). Universality results for "non-standard" architectures appears in (Yarotsky, 2018; Sannai et al., 2019). In addition to proving universality of networks using polynomial layer, Yarotsky (2018) also prove universality for standard MLPs equivariant to Abelian groups. A similar results follows as a corollary to our main theorem.

A parallel line of work in equivariant deep learning studies linear action of a group beyond permutations. The resulting equivariant linear layers can be written using convolution operations (Cohen & Welling, 2016b; Kondor & Trivedi, 2018). When limited to permutation groups, group convolution is simply another expression of parameter-sharing (Ravanbakhsh et al., 2017b); see also Section 2.3. However, in working with linear representations, one may move beyond finite groups (Cohen et al., 2019a; Kondor & Trivedi, 2018); see also (Wood & Shawe-Taylor, 1996). Some applications include equivariance to isometries of the Euclidean space (Weiler & Cesa, 2019; Worrall et al., 2017), and sphere (Cohen et al., 2018). Extension of this view to manifolds is proposed in (Cohen et al., 2019b). Finally, a third line of work in equivariant deep learning that involves a

[1]School of Computer Science, McGill University, Montreal Canada. [2]Mila - Quebec AI Institute.. Correspondence to: Siamak Ravanbakhsh <siamak@cs.mcgill.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

specialized architecture and learning procedure is that of Capsule networks (Sabour et al., 2017; Hinton et al., 2018); see (Lenssen et al., 2018) for a group theoretic generalization.

2. Preliminaries

Let $\mathcal{G} = \{g\}$ be a finite group. We define the *action* of this group on two finite sets $\mathbb{N}$ and $\mathbb{M}$ of input and output units in a feedforward layer. Using these actions which define permutation groups we then define equivariance and invariance. In detail, $\mathcal{G}$-action on the set $\mathbb{N}$ is a structure preserving map (homomorphism) $a : \mathcal{G} \to \mathcal{S}_\mathbb{N}$, into the symmetric group $\mathcal{S}_\mathbb{N}$, the group of all permutations of $\mathbb{N}$. The *image* of this map is a permutation group $\mathcal{G}_\mathbb{N} \leq \mathcal{S}_\mathbb{N}$. Instead of writing $[a(g)](n)$ for $g \in \mathcal{G}$ and $n \in \mathbb{N}$, we use the short notation $g \cdot n = g^{-1}n$ to denote this action. Let $\mathbb{M}$ be another $\mathcal{G}$-set, where the corresponding permutation action $\mathcal{G}_\mathbb{M} \leq \mathcal{S}_\mathbb{M}$ is defined by $b : \mathcal{G} \to \mathcal{S}_\mathbb{M}$. $\mathcal{G}$-action on $\mathbb{N}$ naturally extends to $\mathbf{x} \in \mathbb{R}^\mathbb{N}$ by $g \cdot \mathbf{x}_n \doteq \mathbf{x}_{g \cdot n}\ \forall g \in \mathcal{G}_\mathbb{N}$. More conveniently, we also write this action as $\mathbf{A}_g\mathbf{x}$, where $\mathbf{A}_g$ is the permutation matrix form of $a(g, \cdot) : \mathbb{N} \to \mathbb{N}$.

2.1. Invariant and Equivariant Linear Maps

Let the real matrix $\mathbf{W} \in \mathbb{R}^{|\mathbb{N}|\times|\mathbb{M}|}$ denote a linear map $\mathbf{W} : \mathbb{R}^{|\mathbb{N}|} \to \mathbb{R}^{|\mathbb{M}|}$. We say this map is $\mathcal{G}$-equivariant iff

$$\mathbf{B}_g\mathbf{W}\mathbf{x} = \mathbf{W}\ \mathbf{A}_g\mathbf{x} \quad \forall \mathbf{x} \in \mathbb{R}^\mathbb{N}, g \in \mathcal{G}. \tag{1}$$

where similar to $\mathbf{A}_g$, the permutation matrix $\mathbf{B}_g$ is defined based on the action $b(\cdot, g) : \mathbb{M} \to \mathbb{M}$. In this definition, we assume that the group action on the input is *faithful* – that is a is injective, or $\mathcal{G}_\mathbb{N} \cong \mathcal{G}$. If the action on the output index set $\mathbb{M}$ is not faithful, then the *kernel* of this action is a non-trivial *normal subgroup* of $\mathcal{G}$, $\ker(b) \triangleleft \mathcal{G}$. In this case $\mathcal{G}_\mathbb{M} \cong \mathcal{G}/\ker(b)$ is a *quotient group*, and it is more accurate to say that $\mathbf{W}$ is invariant to $\ker(b)$ and equivariant to $\mathcal{G}/\ker(b)$. Using this convention $\mathcal{G}$-equivariance and $\mathcal{G}$-invariance correspond to extreme cases of $\ker(b) = \mathcal{G}$ and $\ker(b) = \{e\}$. Moreover, composition of such invariant-equivariant functions preserves this property, motivating design of deep networks by stacking equivariant layers.

2.2. Orbits and Homogeneous Spaces

$\mathcal{G}_\mathbb{N}$ partitions $\mathbb{N}$ into *orbits* $\mathbb{N}_1, \dots, \mathbb{N}_O$, where $\mathcal{G}_\mathbb{N}$ is *transitive* on each orbit, meaning that for each pair $n_1, n_2 \in \mathbb{N}_o$, there is at least one $g \in \mathcal{G}_\mathbb{N}$ such that $g \cdot n_1 = n_2$. If $\mathcal{G}_\mathbb{N}$ has a single orbit, it is transitive, and $\mathbb{N}$ is called a *homogeneous space* for $\mathcal{G}$. If moreover the choice of $g \in \mathcal{G}_\mathbb{N}$ with $g \cdot n_1 = n_2$ is unique, then $\mathcal{G}_\mathbb{N}$ is called *regular*.

Given a subgroup $\mathcal{H} \leq \mathcal{G}$ and $g \in \mathcal{G}$, the *right coset* of $\mathcal{H}$ in $\mathcal{G}$, defined as $\mathcal{H}g \doteq \{hg, h \in \mathcal{H}\}$ is a subset of $\mathcal{G}$. For a fixed $\mathcal{H} \leq \mathcal{G}$, the set of these right-cosets, $\mathcal{H}\backslash\mathcal{G} = \{\mathcal{H}g, g \in \mathcal{G}\}$, form a partition of $\mathcal{G}$. $\mathcal{G}$ naturally acts on the right coset space, where $g' \cdot (\mathcal{H}g) \doteq \mathcal{H}(gg')$ sends one coset to another. The significance of this action is that "any" transitive $\mathcal{G}$-action is isomorphic to $\mathcal{G}$-action on some right coset space. To see why, note that in this action any $h \in \mathcal{H}$ *stabilizes* the coset $\mathcal{H}e$, because $h \cdot \mathcal{H}e = \mathcal{H}e$.[1] Therefore in any action the stabilizer identifies the coset space.

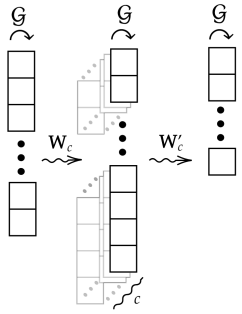

Figure 1. The equivariant MLP of (16). The symbol $\curvearrowright$ indicates $\mathcal{G}$-action on the units, $\mathbf{W}_c$ and $\mathbf{W}'_c$ for all channels of the hidden layer $c = 1, \dots, C$ are constrained by the parameter-sharing of (3). If $\mathcal{G}$-action on the hidden layer is regular, the number of channels can grow to approximate any continuous $\mathcal{G}$-equivariant function with an arbitrary accuracy. Bias terms are not shown.

2.3. Parameter-Sharing and Group-CNNs View

Consider the equivariance condition of (1). Since the equality holds for all $\mathbf{x} \in \mathbb{R}^\mathbb{N}$, and using the fact that the inverse of a permutation matrix is its transpose, the equivariance constraint reduces to

$$\mathbf{B}_g\mathbf{W}\mathbf{A}_g^\top = \mathbf{W} \quad \forall g \in \mathcal{G}. \tag{2}$$

The equation above ties the parameters within the orbits of $\mathcal{G}$-action on rows and columns of $\mathbf{W}$:

$$\mathbf{W}(m, n) = \mathbf{W}(g \cdot m, g \cdot n) \forall g \in \mathcal{G}, n, m \in \mathbb{N} \times \mathbb{M} \tag{3}$$

where $\mathbf{W}(g \cdot m, g \cdot n)$ is an element of the matrix $\mathbf{W}$. This type of group action on Cartesian product space is sometimes called the *diagonal* action. In this case, the action is on the Cartesian product of rows and columns of $\mathbf{W}$.

We saw that any homogenous $\mathcal{G}$-space is isomorphic to a coset space. Using $\mathbb{N} \cong \mathcal{H}\backslash\mathcal{G}$ and $\mathbb{M} \cong \mathcal{K}\backslash\mathcal{G}$, the

[1] More generally, when $\mathcal{G}$ acts on the coset $\mathcal{H}a \in \mathcal{H}\backslash\mathcal{G}$, all $g \in a^{-1}\mathcal{H}a$ stabilize $\mathcal{H}a$. Since $g = a^{-1}ha$ for some $h \in \mathcal{H}$, we have $(a^{-1}ha) \cdot \mathcal{H}a = \mathcal{H}(aa^{-1}ha) = \mathcal{H}a$. This means that any transitive $\mathcal{G}$-action on a set $\mathbb{N}$ may be identified with the stabilizer subgroup $\mathcal{G}_n \doteq \{g \in \mathcal{G}\ s.t.\ g \cdot n = n\}$, for a choice of $n \in \mathbb{N}$. This gives a bijection between $\mathbb{N}$ and the right coset space $\mathcal{G}_n\backslash\mathcal{G}$.

parameter-sharing constraint of (2) becomes

$$\mathbf{W}(\mathcal{K}g, \mathcal{H}g') = \mathbf{W}(g^{-1} \cdot \mathcal{K}g, g^{-1} \cdot \mathcal{H}g') \quad (4)$$
$$= \mathbf{W}(\mathcal{K}, \mathcal{H}g'g^{-1}) \forall g, g' \in \mathcal{G}, \quad (5)$$

Since we can always multiply both indices to have the coset $\mathcal{K}$ as the first argument, we can replace the matrix $\mathbf{W}$ with the vector w, such that $\mathbf{W}(\mathcal{K}g, \mathcal{H}g') = w(\mathcal{H}g'g^{-1}) \quad \forall g, g' \in \mathcal{G}$. This rewriting also enables us to express the matrix vector multiplication of the linear map $\mathbf{W}$ in the form of cross-correlation of input and a kernel w

$$[\mathbf{W}\mathbf{x}](n) = [\mathbf{W}\mathbf{x}](\mathcal{K}g) \quad (6)$$
$$= \sum_{\mathcal{H}g' \in \mathcal{H}\backslash\mathcal{G}} \mathbf{W}(\mathcal{K}g, \mathcal{H}g')\mathbf{x}(\mathcal{H}g') \quad (7)$$
$$= \sum_{\mathcal{H}g' \in \mathcal{H}\backslash\mathcal{G}} w(\mathcal{H}g'g^{-1})\mathbf{x}(\mathcal{H}g') \quad (8)$$

This relates the parameter-sharing view of equivariant maps (4) to the convolution view (8). Therefore, the universality results in the following extends to group convolution layers (Cohen & Welling, 2016a; Cohen et al., 2019a), for finite groups.

Equivariant Affine Maps We may extend our definition, and consider *affine* $\mathcal{G}$-maps $\mathbf{W}\mathbf{x} + \mathbf{b}$, by allowing an "invariant" *bias* parameter $\mathbf{b} \in \mathbb{R}^{|\mathbb{M}|}$ satisfying

$$\mathbf{B}_g\mathbf{b} = \mathbf{b}. \quad (9)$$

This implies a parameter sharing constraint $\mathbf{b}(m) = \mathbf{b}(g \cdot m)$. For homogeneous $\mathbb{M}$, this constraint enforces a *scalar* bias. Beyond homogeneous spaces, the number of free parameters in $\mathbf{b}$ grows with the number of orbits.

2.4. Invariant and Equivariant MLPs

One may stack multiple layers of equivariant affine maps with multiple channels, followed by a non-linearity, so as to build an *equivariant MLP*. One layer of this equivariant MLP a.k.a. *equivariant network* is given by:

$$\mathbf{x}_c^{(\ell)} = \sigma\left(\sum_{c'=1}^{C^{(\ell-1)}} \mathbf{W}_{c,c'}^{(\ell)}\mathbf{x}_{c'}^{(\ell-1)} + \mathbf{b}_c^{(\ell)}\right),$$

where $1 \leq c' \leq C^{(\ell-1)}$ and $1 \leq c \leq C^{(\ell)}$ index the input and output channels respectively, $\mathbf{x}^{(\ell)}$ is the output of layer $1 \leq \ell \leq L$, with $\mathbf{x}^{(0)} = \mathbf{x}$ denoting the original input. Here, we assume that $\mathcal{G}$ faithfully acts on all $\mathbf{x}_c^{(\ell)} \in \mathbb{R}^{\mathbb{H}^{(\ell)}} \quad \forall c, \ell$, with $\mathbb{H}^{(0)} = \mathbb{N}$ and $\mathbb{H}^{(L)} = \mathbb{M}$. The parameter matrices $\mathbf{W}^{\ell}_{c^{(\ell)},c^{(\ell)}} \in \mathbb{R}^{\mathbb{H}^{(\ell-1)} \times \mathbb{H}^{(\ell)}}$, and the bias vector $\mathbf{b}_c^{(\ell)} \in \mathbb{R}^{\mathbb{H}^{(\ell)}}$ are constrained by the parameter-sharing conditions (2) and (9) respectively. In an *invariant MLP* the faithfulness condition for $\mathcal{G}$-action on the hidden and output layers are lifted. In practice, it is common to construct invariant networks by first constructing an equivariant network followed by pooling over $\mathbb{H}^{(L)}$.

3. Universality Results

This section presents two new results on universality of both invariant and equivariant networks with a single hidden layer ($L = 2$). Formally, we can claim that a $\mathcal{G}$-equivariant MLP $\hat{\psi} : \mathbb{R}^{|\mathbb{N}|} \to \mathbb{R}^{|\mathbb{M}|}$ is a *universal $\mathcal{G}$-equivariant approximator*, if for any $\mathcal{G}$-equivariant continuous function $\psi : \mathbb{R}^{|\mathbb{N}|} \to \mathbb{R}^{|\mathbb{M}|}$, any compact set $\mathbb{K} \subset \mathbb{R}^{|\mathbb{N}|}$, and $\epsilon > 0$, there exists a choice of parameters, and number of channels such that $||\psi(\mathbf{x}) - \hat{\psi}(\mathbf{x})|| < \epsilon \; \forall \mathbf{x} \in \mathbb{K}$.

Theorem 3.1. *A $\mathcal{G}$-invariant network*

$$\hat{\psi}(\mathbf{x}) = \sum_{c=1}^{C} w'_c \mathbf{1}^\top \sigma\big(\mathbf{W}_c\mathbf{x} + b_c\big). \quad (10)$$

with a single hidden layer, on which $\mathcal{G}$ acts regularly is a universal $\mathcal{G}$-invariant approximator. Here, $\mathbf{1} = \underbrace{[1, \dots, 1]^\top}_{|\mathcal{G}|}$ and , $b_c, w'_c \in \mathbb{R}$.

Proof. The first step follows the symmetrisization argument (Yarotsky, 2018). Since MLP is a universal approximator, for any compact set $\mathbb{K} \subset \mathbb{R}^{|\mathbb{N}|}$, we can find ψ_{MLP} such that for any $\epsilon > 0$, $|\psi(\mathbf{x}) - \psi_{MLP}(\mathbf{x})| \leq \epsilon$ for $\mathbf{x} \in \mathbb{K}$. Let $\mathbb{K}_{sym} = \{\bigcup_{g \in \mathcal{G}} \mathbf{A}_g\mathbf{x} | \mathbf{x} \in \mathbb{K}\}$ denote the symmetrisized $\mathbb{K}$, which is again a compact subset of $\mathbb{R}^{\mathbb{N}}$ for finite $\mathcal{G}$. Let ψ_{MLP+} approximate ψ on the symmetrisized compact set $\mathbb{K}_{sym}$. It is then easy to show that for $\mathcal{G}$-invariant ψ, the *symmetrisized MLP* $\psi_{sym}(\mathbf{x}) = \frac{1}{|\mathcal{G}|}\sum_{g \in \mathcal{G}} \psi_{MLP+}(\mathbf{A}_g\mathbf{x})$ also approximates ψ

$$|\psi(\mathbf{x}) - \psi_{sym}(\mathbf{x})| = |\psi(\mathbf{x}) - \frac{1}{|\mathcal{G}|}\sum_{g \in \mathcal{G}} \psi_{MLP+}(\mathbf{x})| \quad (11)$$
$$\leq \frac{1}{|\mathcal{G}|}\sum_{g \in \mathcal{G}} |\psi(\mathbf{A}_g\mathbf{x}) - \psi_{MLP}(\mathbf{A}_g\mathbf{x})| \leq \epsilon. \quad (12)$$

Next step, is to show that ψ_{sym} is equal to $\hat{\psi}$ of (10), for some parameters $\mathbf{W}_c \in \mathbb{R}^{|\mathbb{H}| \times |\mathbb{N}|}$ constrained so that $\mathbf{H}_g\mathbf{W}_c = \mathbf{W}_c\mathbf{A}_g \forall g \in \mathcal{G}$, where $\mathbf{A}_g$ and $\mathbf{H}_g$ are the permutation representation of $\mathcal{G}$ action on the input and the hidden layer respectively.

$$\psi_{sym}(\mathbf{x}) = \frac{1}{|\mathcal{G}|}\sum_{g \in \mathcal{G}}\sum_{c=1}^{C} w'_c \sigma\big(\mathbf{w}_c^\top(\mathbf{A}_g\mathbf{x})\big) \quad (13)$$

$$= \sum_{c=1}^{C} \frac{w'_c}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \sigma\big((\mathbf{w}_c^\top \mathbf{A}_g)\mathbf{x}\big) \tag{14}$$

$$= \sum_{c=1}^{C} \tilde{w}_c \mathbf{1}^\top \sigma \left(\underbrace{\begin{bmatrix} -\mathbf{w}_c^\top \mathbf{A}_{g_1} - \\ \vdots \\ -\mathbf{w}_c^\top \mathbf{A}_{g_{|\mathbb{H}|}} - \end{bmatrix}}_{\mathbf{W}_c} \mathbf{x} \right). \tag{15}$$

where in the last step we put the summation terms into rows of the matrix $\mathbf{W}_c$, and performed the summation using multiplication by $\mathbf{1}^\top$. $\tilde{w}_c$ is the rescaled w'_c. Since the summation in (13) is over $g \in \mathcal{G}$, each row of $\mathbf{W}_c$ and therefore each hidden unit is "attached" to exactly one group member, which translates to having a *principal homogeneous space*, a.k.a. a regular $\mathcal{G}$-set. Note that we have the freedom to choose the rows to have any order, corresponding to a different order in summation, which means that the choice of a particular principal homogeneous space is irrelevant.

Now we show that the parameter matrix $\mathbf{W}_c \in \mathbb{R}^{|\mathbb{H}|\times|\mathbb{N}|}$ above satisfy the parameter-sharing constraint $\mathbf{W}_c\mathbf{A}_g = \mathbf{H}_g\mathbf{W}_c \; \forall g \in \mathcal{G}$:

$$\mathbf{H}_g\mathbf{W}_c\mathbf{A}_g^{-1} = \begin{bmatrix} \mathbf{w}_c^\top \mathbf{A}_{g_1 g} \\ \vdots \\ \mathbf{w}_c^\top \mathbf{A}_{g_{|\mathbb{H}|} g} \end{bmatrix} \mathbf{A}_{g^{-1}} = \begin{bmatrix} \mathbf{w}_c^\top \mathbf{A}_{g_1} \\ \vdots \\ \mathbf{w}_c^\top \mathbf{A}_{g_{|\mathbb{H}|}} \end{bmatrix} = \mathbf{W}_c$$

where the first equality follows from the fact that row indexed by g_r is moved to the row $g \cdot g_r = g_r g^{-1}$: $\mathbf{H}_g\mathbf{A}_{g_r} = \mathbf{A}_{g\cdot g_r} = \mathbf{A}_{g_r g^{-1}}$. Therefore, the current row $g_{r'}$ was previously $g^{-1}\cdot g_{r'} = g_{r'}g$. The second equality follows from $\mathbf{A}_g^{-1}$ is acting from the right, and no further inversion is needed $\mathbf{A}_{g_r g}\mathbf{A}_g^{-1} = \mathbf{A}_{g_r g g^{-1}} = \mathbf{A}_{g_r}$. This shows that a $\mathcal{G}$-invariant network with a single hidden layer on which $\mathcal{G}$ acts regularly is equivalent to a symmetricized MLP, and therefore for some number of channels, it is a universal approximator of $\mathcal{G}$-invariant functions. □

This result should not be surprising since the size of a regular hidden layer grows with the group, and as it is evident from the proof, *an equivariant MLP with a regular hidden layer implicitly averages the output over all transformations of the input.* Next, we apply a similar idea to prove the universality of the *equivariant* MLPs with a regular hidden layer.

Theorem 3.2. *A $\mathcal{G}$-equivariant MLP*

$$\hat{\psi}(\mathbf{x}) = \sum_{c=1}^{C} \mathbf{W'}_c \, \sigma\big(\mathbf{W}_c\mathbf{x} + b_c\big). \tag{16}$$

with a single regular *hidden layer is a universal $\mathcal{G}$-equivariant approximator.*

Proof. In this setting, symmetricization, using the so-called *Reynolds operator* (Sturmfels, 2008), for the universal MLP is given by

$$\psi_{sym}(\mathbf{x}) = \frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \mathbf{B}_{g^{-1}} \sum_{c=1}^{C} \mathbf{w}'_c \sigma\left(\mathbf{w}_c^\top \mathbf{A}_g \mathbf{x} + b_c\right) \tag{17}$$

where $\mathbf{w}_c \in \mathbb{R}^{|\mathbb{N}|}$ and $\mathbf{w}'_c \in \mathbb{R}^{\mathbb{M}}$ are the weight vectors in the first and second layer associated with hidden unit c. Our objective is to show that this symmetrisized MLP is equivalent to the equivariant network of (16), in which $\mathbf{W}'_c \in \mathbb{R}^{|\mathbb{M}|\times|\mathbb{H}|}$, and $\mathbf{W}_c \in \mathbb{R}^{|\mathbb{H}|\times|\mathbb{N}|}$ use parameter-sharing to satisfy

$$\mathbf{H}_g\mathbf{W}_c = \mathbf{W}_c\mathbf{A}_g \text{ and } \mathbf{B}_g\mathbf{W}'_c = \mathbf{W}'_c\mathbf{H}_g \; \forall g \in \mathcal{G}. \tag{18}$$

Here, $\mathbf{A}_g$, $\mathbf{B}_g$ and $\mathbf{H}_g$ are the permutation representations of $\mathcal{G}$ action on the input, the output, and the hidden layer respectively.

First, rewrite the symmetrisized MLP as

$$\psi_{sym}(\mathbf{x}) = \sum_{c=1}^{C} \sum_{g\in\mathcal{G}} \mathbf{B}_{g^{-1}} \mathbf{w}'_c \sigma\big(\mathbf{w}_c^\top \mathbf{A}_g \mathbf{x} + b_c\big)$$

$$= \sum_{c=1}^{C} \mathbf{W}'_c \sigma\big(\mathbf{W}_c\mathbf{x}\big)$$

$$\text{where} \quad \mathbf{W}'_c = \begin{bmatrix} | & & | \\ \mathbf{B}_{g_1^{-1}}\mathbf{w}'_c & \dots & \mathbf{B}_{g_{|\mathcal{G}|}^{-1}}\mathbf{w}'_c \\ | & & | \end{bmatrix}$$

$$\mathbf{W}_c = \begin{bmatrix} -\mathbf{w}_c\mathbf{A}_{g_1} - \\ \vdots \\ -\mathbf{w}_c\mathbf{A}_{g_{|\mathcal{G}|}} - \end{bmatrix},$$

and the $\frac{1}{|\mathcal{G}|}$ factor is absorbed in one of the weights. It remains to show that the two matrices above satisfy the equivariance condition $\mathbf{H}_g\mathbf{W}_c = \mathbf{W}_c\mathbf{A}_g$ and $\mathbf{B}_g\mathbf{W}'_c = \mathbf{W}'_c\mathbf{H}_g$. The proof for $\mathbf{W}_c$ is identical to the invariant network case.

For $\mathbf{W}'_c$, we use a similar approach.

$$\mathbf{B}_g\mathbf{W}'_c\mathbf{H}_g^{-1} = \begin{bmatrix} | & & | \\ \mathbf{B}_g\mathbf{B}_{g_1^{-1}g}\mathbf{w}'_c & \dots & \mathbf{B}_g\mathbf{B}_{g_{|\mathcal{G}|}^{-1}g}\mathbf{w}'_c \\ | & & | \end{bmatrix}$$

$$= \begin{bmatrix} | & & | \\ \mathbf{B}_{g_1^{-1}}\mathbf{w}'_c & \dots & \mathbf{B}_{g_{|\mathcal{G}|}^{-1}}\mathbf{w}'_c \\ | & & | \end{bmatrix} = \mathbf{W}'_c.$$

In the first step, since $\mathbf{H}_g^{-1} = \mathbf{H}_{g^{-1}}$ is acting on the right, it moves the column indexed by g_l^{-1} to $g_l^{-1}g^{-1}$. This means

that the column currently at $g_{l'}^{-1}$ is $g_{l'}^{-1}g$. The second step uses the following: $\mathbf{B}_g\mathbf{B}_{g_l^{-1}g} = \mathbf{B}_{g\cdot(g_l^{-1}g)} = \mathbf{B}_{g_l^{-1}gg^{-1}} = \mathbf{B}_{g_l^{-1}}$. This, proves the equality of the symmetrisize MLP (17) to the equivariant MLP of (16). However, a similar argument to the proof of invariant case, shows the universality of ψ_{sym}. Putting these together, completes the proof of Theorem 3.2. □

3.1. Universality for Abelian Groups

In the case where $\mathcal{G}$ is an *Abelian group*, any faithful transitive action is regular, meaning that the hidden layer in a $\mathcal{G}$-equivariant neural network is necessarily regular. Combined with Theorem 3.2, this leads to an unconditional universality result for Abelian groups. A similar result for Abelian groups appears in (Yarotsky, 2018).

Corollary 1. *For Abelian group $\mathcal{G}$, a $\mathcal{G}$-equivariant (invariant) neural network with a single hidden layer is a universal approximator of continuous $\mathcal{G}$-equivariant (invariant) functions on compact subsets of $\mathbb{R}^{|\mathbb{N}|}$.*

A corollary to this is the universality of a Convolutional Neural Network (CNN) with a single hidden layer.

Corollary 2 (Universality of CNNs)**.** *For an arbitrary input-output dimensions, a CNN with a single hidden layer, full kernels, and cyclic padding is a universal approximator of continuous circular translation equivariant (invariant) functions.*

Use of the term circular, both in padding and translation is because of the need to work with finite translations, which are produce as the result of the action of a product of cyclic groups.[2]

3.2. Universality for Regular Steerable CNNs

In building models equivariant to subgroups of Euclidean isometries (translation, rotation and reflection), a simple solution is to consider the action of (circular) translation group on rotated and/or reflected copies of each feature-map (Dieleman et al., 2016). This approach is formalized in (Cohen & Welling, 2016b) for general semi-direct product $\mathcal{G} = \mathcal{N} \rtimes \mathcal{H}$ and linear representations. In the setting where the action of $\mathcal{N}$ and $\mathcal{H}$ on the fibers and the base-space are both regular, $\mathcal{G}$-action is also regular, and as a corollary to Theorem 3.2 steerable CNN becomes universal. For example, this is the case in the practical setup where the feature-maps are rotated and reflected.

[2]Input can be zero-padded, before circular padding, so that Corollary 2 guarantees universal approximation of translation equivariant functions, where translations are bounded by the size the original input.

3.3. Universality for High-Order Hidden Layers

$\mathcal{G}$-action on the hidden units $\mathbb{H}$ naturally extends to its simultaneous action on the Cartesian product $\mathbb{H}^D = \mathbb{H} \times \ldots \times \mathbb{H}$:

$$g \cdot (h_1, \ldots, h_D) \doteq (g \cdot h_1, \ldots, g \cdot h_D).$$

We call this an *order D product space*. Product spaces are used in building high-order layers in $\mathcal{G}$-equivariant networks in several recent works (Kondor et al., 2018; Maron et al., 2018; Keriven & Peyré, 2019; Albooyeh et al., 2019). Maron et al. (2019) show that for

$$D \geq \frac{1}{2}|\mathbb{H}|\ (|\mathbb{H}| - 1), \tag{19}$$

such MLPs with multiple hidden layers of order D become universal $\mathcal{G}$-*invariant* approximators. In this section, we show that better bounds for D that guarantees universal invariance and equivariance follows from the universality results of Theorems 3.1 and 3.2. The next section provides an in-depth analysis of product spaces that not only gives an alternative proof of the theorems below, but also could lead to yet better bounds.[3]

Theorem 3.3. *Let $\mathcal{G}$ act faithfully on $\mathbb{H} \cong [\mathcal{H}\backslash\mathcal{G}]$. Then $\mathbb{H}^D$ has a regular orbit for any*

$$D \geq \log_2(|\mathcal{H}|)$$

and therefore, by Theorem 3.2, an order D hidden layer guarantees universal equivariance.

Proof. If $\mathcal{G}$ acts faithfully on $\mathbb{H}$, the intersection of the stabilisers of all the points in $\mathbb{H}$ is trivial – *i.e.*, $\mathrm{Core}_{\mathcal{G}}(\mathcal{H}) = \{e\}$. If instead of taking the intersection of the stabilisers of all $h \in \mathbb{H}$, we can just take the intersection of the stabilisers of D (carefully chosen) points, we will know there is a regular orbit in $\mathbb{H}^D$. That is because the stabiliser of a point in $\mathbb{H}^d$ is the intersection of the stabilisers of its elements in $\mathbb{H}$, that is $\mathrm{Stab}_{\mathcal{G}}(h_1, ..., h_D) = \bigcap_{d=1}^{D} \mathrm{Stab}_{\mathcal{G}}(h_d)$. So the question is for what value of D can we find D points such that the intersection of their stabilisers is trivial. We work recursively to find a bound on D.

Start with just one point $h_1 in \mathbb{H}^1$, and assume its stabiliser is of size s_1. Now assume we have a point $(h_1, ..., h_d)$ in $\mathbb{H}^d$

[3]The beautiful proof for the following theorem was proposed by an anonymous reviewer. The original proof uses the ideas discussed in the next section and appears later in the paper.

such that its stabiliser is of size s_d. If $s_d = 1$, we are done. Otherwise, since the action is faithful, there has to exist a point h_{d+1} such that the intersection of all the stabilisers of $h_1, ..., h_{d+1}$ is a strictly smaller subgroup of the stabiliser of $(h_1, ..., h_d)$. The size of a proper subgroup is at most half the size of the original group and therefore $s_{d+1} < s_d/2$. Therefore, for each additional point the size of stabilizer at least half of the previous stabilizer. It follows that for any $D \geq \log_2(|\mathcal{H}|)$, $[\mathcal{H}\backslash\mathcal{G}]^D = \mathbb{H}^D$ has an orbit with a trivial stabilizer. □

Since the largest stabilizer for any action on $\mathbb{H}$ is $\mathcal{S}_{|\mathbb{H}|-1}$, we can use a lower-bound for D, in Theorem 3.3 that is independent of the stabilizer sub-group $\mathcal{H}$. The following bound follows from the Sterling's approximation $N! < N^{N+\frac{1}{2}}e^{-N+1}$ to the size of the largest possible stabilizer $|\mathcal{S}_{|\mathbb{H}|-1}| = (|\mathbb{H}-1|)!$.

Corollary 3. *The high-order $\mathcal{G}$-set of hidden units $\mathbb{H}^D$, with $N = |\mathbb{H}|$ has a regular orbit for*

$$D \geq \lceil (N - \frac{1}{2}) \log_2(N-1) - (N-2)\log_2(e) \rceil$$

and following Theorem 3.2 the corresponding equivariant MLP is universal approximator of continuous $\mathcal{G}$-equivariant functions.

4. Decomposition of Product $\mathcal{G}$-Sets

A prerequisite to analysis of product $\mathcal{G}$-sets is their classification, which also leads to classification of all $\mathcal{G}$-maps based on their input/output $\mathcal{G}$-sets.

4.1. Classification of $\mathcal{G}$-Sets and $\mathcal{G}$-Maps

Recall that any transitive $\mathcal{G}$-set $\mathbb{N}$ is isomorphic to a right-coset space $\mathcal{H}\backslash\mathcal{G}$. However, the right cosets $\mathcal{H}\backslash\mathcal{G}$ and $(g^{-1}\mathcal{H}g)\backslash\mathcal{G} \quad \forall g \in \mathcal{G}$ are themselves isomorphic. [4] This also means what we care about is *conjgacy classes of subgroups* $[\mathcal{H}] = \{g^{-1}\mathcal{H}g \mid g \in \mathcal{G}\}$, which classifies right-coset spaces up to conjugacy $[\mathcal{H}\backslash\mathcal{G}] = \{g^{-1}\mathcal{H}g\backslash\mathcal{G} \mid g \in \mathcal{G}\}$. We used the bracket to identify the conjugacy class. In this notation, for $\mathcal{H}, \mathcal{H}' \leq \mathcal{G}$, we say $[\mathcal{H}] < [\mathcal{H}']$, iff $g^{-1}\mathcal{H}g < \mathcal{H}'$, for some $g \in \mathcal{G}$.

A $\mathcal{G}$-set is transitive on each of its orbits, and we can identify each orbit with its stabilizer subgroup. Therefore a list of these subgroups along with their multiplicities completely defines a $\mathcal{G}$-set up to an isomorphism (Rotman, 2012):

$$\mathbb{N} \cong \bigcup_{[\mathcal{H}_i] \leq \mathcal{G}} p_i[\mathcal{H}_i\backslash\mathcal{G}], \tag{20}$$

where $p_1, \ldots, p_I \in \mathbb{Z}^{\geq 0}$ denotes the multiplicity of a right-coset space, and $\mathbb{N}$ has $\sum_{i=1}^{I} p_i$ orbits.

To ensure a faithful $\mathcal{G}$-action on $\mathbb{N}$, a necessary and sufficient condition is for the point-stabilizers $\mathcal{G}_n \forall n \in \mathbb{N}$ to have a trivial intersection. The point-stabilizers within each orbit are conjugate to each other and their intersection which is the largest normal subgroup of $\mathcal{G}$ contained in $\mathcal{H}_i$, is called the *core* of $\mathcal{G}$-action on $[\mathcal{H}_i\backslash\mathcal{G}]$:

$$\mathrm{Core}_{\mathcal{G}}(\mathcal{H}_i) \doteq \bigcap_{g \in \mathcal{G}} g^{-1}\mathcal{H}_i g. \tag{21}$$

Next, we extend the classification of $\mathcal{G}$-sets to $\mathcal{G}$-equivariant maps, a.k.a. $\mathcal{G}$-maps $\mathbf{W} : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{M}}$, by jointly classifying the input and the output index sets $\mathbb{N}$ and $\mathbb{M}$. We may consider a similar expression to (20) for the output index set $\mathbb{M} = \bigcup_{[\mathcal{K}_j] \leq \mathcal{G}} q_j[\mathcal{K}_j\backslash\mathcal{G}]$. The linear $\mathcal{G}$-map $\mathbf{W} : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{M}}$ is then equivariant to $\mathcal{G}/\mathcal{K}$ and invariant to $\mathcal{K} \triangleleft \mathcal{G}$ iff

$$\bigcap_{p_i > 0} \mathrm{Core}_{\mathcal{G}}(\mathcal{H}_i) = \{e\} \text{ and } \bigcap_{q_i > 0} \mathrm{Core}_{\mathcal{G}}(\mathcal{K}_i) = \mathcal{K} \tag{22}$$

where the second condition translates to $\mathcal{K}$ invariance of $\mathcal{G}$-action on $\mathbb{M}$. Note that the first condition is simply ensuring the faithfulness of $\mathcal{G}$-action on $\mathbb{N}$. This result means that the multiplicities $(p_1, \ldots, p_I)$ and $(q_1, \ldots, q_J)$ completely identify a (linear) $\mathcal{G}$-map $\mathbf{W} : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{M}}$ that equivariant to $\mathcal{G}/\mathcal{K}$ and invariant to $\mathcal{K} \triangleleft \mathcal{G}$, up to an isomorphism.

4.2. Diagonal Action on Cartesian Product of $\mathcal{G}$-sets

Previously we classified all $\mathcal{G}$-sets as the disjoint union of homogeneous spaces $\bigcup_{i=1}^{I} p_i[\mathcal{G}_i\backslash\mathcal{G}]$, where $\mathcal{G}$ acts transitively on each orbit. However, as we saw earlier $\mathcal{G}$ also naturally acts on the *Cartesian product* of homogeneous $\mathcal{G}$-sets:

$$\mathbb{N}_1 \times \ldots \times \mathbb{N}_D = (\mathcal{G}_1\backslash\mathcal{G}) \times \ldots \times (\mathcal{G}_D\backslash\mathcal{G})$$

where the action is defined by

$$g \cdot (\mathcal{G}_1 h_1, \ldots, \mathcal{G}_D h_D) \doteq (\mathcal{G}_1(h_1 g), \ldots, \mathcal{G}_D(h_D g)).$$

A special case is when we consider the repeated self-product of the same homogeneous space $\mathbb{H} \cong [\mathcal{H}\backslash\mathcal{G}]$, which as we saw gives an *order D product space*.

$$\mathbb{H}^D \cong [\mathcal{H}\backslash\mathcal{G}]^D = \underbrace{[\mathcal{H}\backslash\mathcal{G}] \times \ldots \times [\mathcal{H}\backslash\mathcal{G}]}_{D \text{ times}}$$

We call this an *order D product space*. The following discussion shows how the product space decomposes into orbits, where the existence of a regular orbit leads to universality.

[4] The stabilizer subgroups of two points in a homogeneous space are conjugate, and therefore $\mathcal{G}$-sets resulting from conjugate choice of right-cosets are isomorphic. To see why stabilizers are conjugate, assume $n = a^{-1} \cdot n$, and $h \in \mathcal{G}_n$, then $aha^{-1} \cdot n = n^{hd} = n^a = n$. Therefore, $a^{-1}ha \in \mathcal{G}_n$. Since conjugation is a bijection, this means $\mathcal{G}_n = a^{-1}\mathcal{G}_n a$.

4.3. Burnside Ring and Decomposition of $\mathcal{G}$-sets

Since *any* $\mathcal{G}$-set can be written as a disjoint union of homogeneous spaces (20), we expect a decomposition of the product $\mathcal{G}$-space in the form

$$[\mathcal{G}_i\backslash\mathcal{G}] \times [\mathcal{G}_j\backslash\mathcal{G}] = \bigcup_{[\mathcal{G}_\ell]\leq\mathcal{G}} \delta^\ell_{i,j}[\mathcal{G}_\ell\backslash\mathcal{G}] \tag{23}$$

Indeed, this decomposition exists, and the multiplicities $\delta^\ell_{i,j} \in \mathbb{Z}^{>0}$, are called the *structure coefficient* of the *Burnside Ring*. The (commutative semi)ring structure is due to the fact that the set of non-isomorphic $\mathcal{G}$-sets $\Omega(\mathcal{G}) = \{\bigcup_{[\mathcal{G}_i]\leq\mathcal{G}} p_i[\mathcal{G}_i\backslash\mathcal{G}] \mid p_i \in \mathbb{Z}^{\geq 0}\}$, is equipped with: 1) a commutative product operation that is the Cartesian product of $\mathcal{G}$-spaces, and; 2) a summation operation that is the disjoint union of $\mathcal{G}$-spaces (Dieck, 2006). A key to analysis of product $\mathcal{G}$-spaces is finding the structure coefficients in (23).

Example 1 (PRODUCT OF SETS)**.** *The symmetric group $\mathcal{S}_\mathbb{N}$ acts faithfully on $\mathbb{N}$, where the stabilizer is $\mathcal{S}_n = \mathcal{S}_{\mathbb{N}-\{n\}}$ – that is the stabilizer of $n \in \mathbb{N}$ is the set of all permutations of the remaining items $\mathbb{N}-\{n\}$. This means $\mathbb{N} \cong [\mathcal{S}_{\mathbb{N}-\{n\}}\backslash\mathcal{S}_\mathbb{N}]$.*

The diagonal $\mathcal{S}_\mathbb{N}$ action on the product space $\mathbb{N}^D$, decomposes into $\sum_i p_i = \mathrm{Bell}(D)$ orbits, where the Bell number is the number of different partitions of a set of D labelled objects (Maron et al., 2018). One may further refine these orbits by their type in the form of (23):

$$[\mathcal{S}_{\mathbb{N}-n}\backslash\mathcal{S}_\mathbb{N}]^D = \bigcup_{d=1}^{D} \mathrm{S}(D,d)[\mathcal{S}_{\mathbb{N}-\{n_1,\ldots,n_d\}}\backslash\mathcal{S}_\mathbb{N}] \tag{24}$$

where the "structure coefficient" $\mathrm{S}(D,d)$ is the Stirling number of the second kind*, and it counts the number of ways D could be partitioned into d non-empty sets. For example, when $D = 2$, one may think of the index set $\mathbb{N}\times\mathbb{N}$ as indexing some $|\mathbb{N}|\times|\mathbb{N}|$ matrix. This matrix decomposes into one ($\mathrm{S}(2,1) = 1$) diagonal $[\mathcal{S}_{\mathbb{N}-\{n\}}\backslash\mathcal{S}_\mathbb{N}]$ and one $\mathrm{S}(2,2) = 1$ set of off-diagonals $[\mathcal{S}_{\mathbb{N}-\{n_1,n_2\}}\backslash\mathcal{S}_\mathbb{N}]$. This decomposition is presented in (Albooyeh et al., 2019), where it is shown that these orbits correspond to "hyper-diagonals" for higher order tensors. For general groups, inferring the structural coefficients is more challenging, as we see shortly.*

From (24) in the example above it follows that an order $D = |\mathbb{N}|$ product of sets contains a regular orbit. The following is a corollary that combines this with the universality results of Theorems 3.1 and 3.2.

Corollary 4. *[Universality of Equivariant Hyper-Graph Networks] A $\mathcal{S}_\mathbb{N}$ equivariant network with a hidden layer of order $D \geq |\mathbb{N}|$, is a universal approximator of $\mathcal{S}_\mathbb{N}$-equivariant (invariant) functions, where the input and output layer may be of any order.*

Note how using group specific analysis gives a better bound of $D \geq N$ compared to group agnostic bound $D \geq N\log(N)$ of Corollary 3. A universality result for the invariant case only, using a quadratic order appears in (Maron et al., 2019), where the MLP is called a *hyper-graph network*. Keriven & Peyré (2019) prove universality for the equivariant case, without giving a bound on the order of the hidden layer, and assuming an output $\mathbb{M} = \mathbb{H}^1$ of degree $D = 1$. In comparison, Corollary 4 uses a linear bound and applies to a much more general setting of arbitrary orders for the input and output product sets. In fact, the universality result is true for arbitrary input-output $\mathcal{S}_\mathbb{N}$-sets.

Linear $\mathcal{G}$-Map as a Product Space For finite groups, the linear $\mathcal{G}$-map $\mathbf{W} : \mathbb{R}^\mathbb{N} \to \mathbb{R}^\mathbb{M}$ is indexed by $\mathbb{M}\times\mathbb{N}$, and therefore it is a product space. In fact the parameter-sharing of (3) ties all the parameters $\mathbf{W}(m,n)$ that are in the same orbit. Therefore, the decomposition (23) also identifies parameter-sharing pattern of $\mathbf{W}$.[5]

Example 2 (EQUIVARIANT MAPS BETWEEN SET PRODUCTS)**.** *Equation (24) gives a closed form for the decomposition of $\mathbb{N}^D$ into orbits. Assuming a similar decomposition for $\mathbb{M}^{D'}$, the equivariant map $\mathbf{W} : \mathbb{R}^{\mathbb{N}^D} \to \mathbb{R}^{\mathbb{M}^{D'}}$ is decomposed in to $\mathrm{Bell}(D+D')$ linear maps corresponding to the orbits of $\mathbb{M}^{D'}\times\mathbb{N}^D$.*

4.3.1. BURNSIDE'S TABLE OF MARKS

Burnside's table of marks simplifies working with the multiplication operation of the Burnside ring, and enables the analysis of $\mathcal{G}$-action on product spaces (Burnside, 1911; Pfeiffer, 1997). The *mark* of $\mathcal{H} \leq \mathcal{G}$ on a finite $\mathcal{G}$-set $\mathbb{N}$, is defined as the number of points in $\mathbb{N}$ fixed by all $h \in \mathcal{H}$:

$$m_\mathbb{N}(\mathcal{H}) \doteq |\{n \in \mathbb{N} \mid h\cdot n = n \;\forall h \in \mathcal{H}\}|. \tag{25}$$

The interesting quality of the number of fixed points is that the total number of fixed points adds up when we add two spaces $\mathbb{N}_1 \cup \mathbb{N}_2$. Also, when considering product spaces $\mathbb{N}_1 \times \mathbb{N}_2$, any combination of points fixed in both spaces

[5] When $\mathbb{N}$ and $\mathbb{M}$ are homogeneous spaces, another characterization the orbits of the product space $[\mathcal{G}_n\backslash\mathcal{G}] \times [\mathcal{G}_m\backslash\mathcal{G}]$ is by showing their one-to-one correspondence with double-cosets $\mathcal{G}_n\backslash\mathcal{G}/\mathcal{G}_m = \{\mathcal{G}_n g\mathcal{G}_m \mid g \in \mathcal{G}\}$.

Table 1. Table of marks $\mathbf{M}_{\mathcal{G}}$.

	$\{e\}$	…	$\mathcal{G}_i$	…	$\mathcal{G}_j$	…	$\mathcal{G}$
$\{e\}\backslash\mathcal{G}$	$\lvert\mathcal{G}\rvert$						
⋮	⋮	⋱					
$\mathcal{G}_i\backslash\mathcal{G}$	$\lvert\mathcal{G}:\mathcal{G}_i\rvert$	…	$\lvert\mathcal{G}:\mathrm{N}_{\mathcal{G}}(\mathcal{G}_i)\rvert$				
⋮	⋮		⋮	⋱			
$\mathcal{G}_j\backslash\mathcal{G}$	$\lvert\mathcal{G}:\mathcal{G}_j\rvert$	…	$m_{\mathcal{G}_j\backslash\mathcal{G}}(\mathcal{G}_i)$	…	$\lvert\mathcal{G}:\mathrm{N}_{\mathcal{G}}(\mathcal{G}_j)\rvert$		
⋮	⋮		⋮		⋮	⋱	
$\mathcal{G}\backslash\mathcal{G}$	1	…	1	…	1	…	1

will be fixed by $\mathcal{H}$. This means

$$m_{\mathbb{N}_1\cup\mathbb{N}_2}(\mathcal{G}i) = m_{\mathbb{N}_1}(\mathcal{G}i) + m_{\mathbb{N}_2}(\mathcal{G}i) \tag{26}$$

$$m_{\mathbb{N}_1\times\mathbb{N}_2}(\mathcal{G}i) = m_{\mathbb{N}_1}(\mathcal{G}i)\, m_{\mathbb{N}_2}(\mathcal{G}i). \tag{27}$$

Now define the *vector of marks* $\mathbf{m}_{\mathbb{N}} : \Omega(\mathcal{G}) \to \mathbb{Z}^n$ as

$$\mathbf{m}_{\mathbb{N}} \doteq [m_{\mathbb{N}}(\mathcal{G}_1), \ldots, m_{\mathbb{N}}(\mathcal{G}_I)]$$

where I is the the number of conjugacy classes of subgroups of $\mathcal{G}$, and we have assume a fixed order on $[\mathcal{G}_i] \leq \mathcal{G}$. Due to Eqs. (26) and (27), given $\mathcal{G}$-sets $\mathbb{N}_1, \ldots, \mathbb{N}_D$, we can perform elementwise addition and multiplication on the vector of integers $\mathbf{m}_{\mathbb{N}_1}, ..., \mathbf{m}_{\mathbb{N}_D}$, to obtain the mark of union and product $\mathcal{G}$-sets respectively. Moreover, the special quality of marks, makes this vector an *injective* homeomorphism: we can work backward from the resulting vector of marks and decompose the union/product space into homogeneous spaces. To facilitate calculation of this vector, for any $\mathcal{G}$-set $\mathbb{N}$, one may use the table of marks.

The *table of marks* for a group $\mathcal{G}$, is the square matrix of marks of all subgroups on all right-coset spaces[6] – that is the element i, j of this matrix is:

$$\mathbf{M}_{\mathcal{G}}(i,j) \doteq m_{\mathcal{G}i\backslash\mathcal{G}}(\mathcal{G}_j) \quad \text{or} \quad \mathbf{M}_{\mathcal{G}} \doteq \begin{bmatrix} \mathbf{m}_{\{e\}\backslash\mathcal{G}} \\ \vdots \\ \mathbf{m}_{\mathcal{G}\backslash\mathcal{G}} \end{bmatrix}. \tag{28}$$

The matrix $\mathbf{M}_{\mathcal{G}}$, has valuable information about the subgroup structure of $\mathcal{G}$. For example, $\mathcal{G}_j$'s action on $\mathcal{G}i\backslash\mathcal{G}$ will have a fixed point, iff $[\mathcal{G}_j] \leq [\mathcal{G}i]$. Therefore, the sparsity pattern in the table of marks, reflects the subgroup lattice structure of $\mathcal{G}$, up to conjugacy.[7]

A useful property of $\mathbf{M}_{\mathcal{G}}$ is that we can use it to find the marks $\mathbf{m}_{\mathbb{N}}$ on any $\mathcal{G}$-set $\mathbb{N} = \sum_i p_i[\mathcal{G}i\backslash\mathcal{G}]$ in $\Omega(\mathcal{G})$ using the expression $\mathbf{m}_{\mathbb{N}} = [p_1, \ldots, p_I]^\top \mathbf{M}_{\mathcal{G}}$. Moreover, the structural constants of (23) can be recovered from the table of Marks

$$\delta^{\ell}_{ij} = \sum_l \mathbf{M}_{\mathcal{G}}(i,l)\mathbf{M}_{\mathcal{G}}(j,l)(\mathbf{M}^{-1}_{\mathcal{G}})(l,\ell). \tag{29}$$

[6] $m_{\mathcal{G}i\backslash\mathcal{G}}(\mathcal{G}_j) = m_{\mathcal{G}i\backslash\mathcal{G}}(g\mathcal{G}_j g^{-1})$, and $m_{\mathcal{G}i\backslash\mathcal{G}}(\mathcal{G}_j) = m_{g\mathcal{G}_i g^{-1}\backslash\mathcal{G}}(\mathcal{G}_j) \quad \forall g \in \mathcal{G}$. Therefore, the table of marks' characterization is up to conjugacy.

[7] The sub-group lattice of $\mathcal{G}$ is a partially ordered set in which the order $\mathcal{G}i < \mathcal{G}_j$ is a subgroup relation, and the greatest and least elements are $\mathcal{G}$ and $\{e\}$ respectively. Any $\mathcal{G}$-set is isomorphic to a right-coset space produced by a member of this lattice. However, we only care about this lattice up to a conjugacy relation. This is because as we saw, the right cosets $\mathcal{H}\backslash\mathcal{G}$ and $(g^{-1}\mathcal{H}g)\backslash\mathcal{G} \quad \forall g \in \mathcal{G}$ are isomorphic.

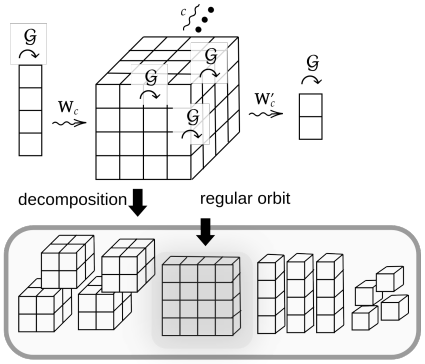

Figure 2. A high-order hidden layer decomposes into orbits, which are characterized by the table of marks. By increasing the order one could guarantee the existence of a regular orbit in the decomposition. By Theorem 3.2 this leads to universal equivariance.

5. Universality of $\mathcal{G}$-Maps on Product Spaces

Using the tools discussed in the previous section, in this section we prove some properties of product spaces that are consequential in design of equivariant maps. Previously we saw that product spaces decompose into orbits, identified by $\delta^{\ell}_{ij} > 0$ in (23). The following theorem states that such product spaces always have orbits that are at least as large as the largest of the input orbits, and at least one of these product orbits is strictly larger than both inputs. For simplicity, this theorem is stated in terms of the stabilizers, rather than the orbits, where by the *orbit-stabilizer* theorem, larger stabilizers correspond to smaller orbits. Also, while the following theorem is stated for the product of homogeneous $\mathcal{G}$-sets, it trivially extends to product of $\mathcal{G}$-sets with multiple orbits.

Theorem 5.1. *Let $[\mathcal{G}i\backslash\mathcal{G}]$ and $[\mathcal{G}_j\backslash\mathcal{G}]$ be transitive $\mathcal{G}$-sets, with $\{e\} < \mathcal{G}i, \mathcal{G}_j < \mathcal{G}$. Their product $\mathcal{G}$-set decomposes into orbits $[\mathcal{G}i\backslash\mathcal{G}] \times [\mathcal{G}_j\backslash\mathcal{G}] = \bigcup_\ell \delta^{\ell}_{ij}[\mathcal{G}_\ell\backslash\mathcal{G}]$, such that:*

(i) *$[\mathcal{G}_\ell] \leq [\mathcal{G}i], [\mathcal{G}_j]$ for all the resulting orbits.*

(ii) *if $\mathcal{G}_j \not\subseteq \mathrm{Core}_{\mathcal{G}}(\mathcal{G}i)$ and $\mathcal{G}i \not\subseteq \mathrm{Core}_{\mathcal{G}}(\mathcal{G}_j)$, then $[\mathcal{G}_\ell] < [\mathcal{G}i], [\mathcal{G}_j]$ for at least one of the resulting orbit.*

Proof. The proof is by analysis of the table of Marks $\mathbf{M}_{\mathcal{G}}$. The vector of mark for the product space is the element-wise

Table 2. Table of marks for the alternating group $\mathcal{A}_5$.

	$\{e\}$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{K}_4$	$\mathcal{C}_5$	$\mathcal{S}_3$	$\mathcal{D}_{10}$	$\mathcal{A}_4$	$\mathcal{A}_5$
$\{e\}\backslash\mathcal{A}_5$	60								
$\mathcal{C}_2\backslash\mathcal{A}_5$	30	2							
$\mathcal{C}_3\backslash\mathcal{A}_5$	20		2						
$\mathcal{K}_4\backslash\mathcal{A}_5$	15	3		3					
$\mathcal{C}_5\backslash\mathcal{A}_5$	12				2				
$\mathcal{S}_3\backslash\mathcal{A}_5$	10	2	1			1			
$\mathcal{D}_{10}\backslash\mathcal{A}_5$	6	2			1		1		
$\mathcal{A}_4\backslash\mathcal{A}_5$	5	1	2	1				1	
$\mathcal{A}_5\backslash\mathcal{A}_5$	1	1	1	1	1	1	1	1	1

product of vector of marks of the input: $\mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]\times[\mathcal{G}_i\backslash\mathcal{G}]} = \mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]} \odot \mathbf{m}_{[\mathcal{G}_j\backslash\mathcal{G}]}$. The same vector, can be written as a linear combination of rows of $\mathbf{M}_{\mathcal{G}}$, with non-negative integer coefficients: $\mathbf{m}_{\mathcal{G}_i\backslash\mathcal{G}} \odot \mathbf{m}_{\mathcal{G}_j\backslash\mathcal{G}} = \sum_\ell \delta^\ell_{ij} \mathbf{m}_{[\mathcal{G}_\ell\backslash\mathcal{G}]}$. For convenience we assume a *topological ordering* of the conjugacy class of subgroups $\{e\} = \mathcal{G}_1, \ldots, \mathcal{G}_i, \ldots, \mathcal{G}_I = \mathcal{G}$ consistent with their partial order – that is $[\mathcal{G}_i] \not> [\mathcal{G}_j] \forall j > i$. This means that $\mathbf{M}_{\mathcal{G}}$ is lower-triangular, with nonzero diagonals; see Table 1. Three important properties of this table are (Pfeiffer, 1997): (1) the sparsity pattern in $\mathbf{M}_{\mathcal{G}}$ reflects the subgroup relation: $\mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]}(\ell) > 0$ iff $\mathcal{G}_\ell \leq \mathcal{G}_i$. (2) the first column is the index of $\mathcal{G}_i$ in $\mathcal{G}$: $\mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]}(1) = |\mathcal{G} : \mathcal{G}_i| \quad \forall i$. (3) the diagonal element is the index of the normalizer: $\mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]}(i) = |\mathcal{G} : N_{\mathcal{G}}(\mathcal{G}_i)| \ \forall i$, where the *normalizer* of $\mathcal{H}$ in $\mathcal{G}$ is defined as the largest intermediate subgroup of $\mathcal{G}$ in which $\mathcal{H}$ is normal: $N_{\mathcal{G}}(\mathcal{H}) = \{g \in \mathcal{G} \mid g\mathcal{H}g^{-1} = \mathcal{H}\}$.

(i) From (1) it follows that the non-zeros of the product $(\mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]} \odot \mathbf{m}_{[\mathcal{G}_j\backslash\mathcal{G}]})(\ell) > 0$ correspond to $\mathcal{G}_\ell \leq [\mathcal{G}_i]$ and $\mathcal{G}_\ell \leq [\mathcal{G}_j]$. Since the only rows of $\mathbf{M}_{\mathcal{G}}$ with such non-zero elements are $\mathbf{m}_{[\mathcal{G}_\ell\backslash\mathcal{G}]}$ for $\mathcal{G}_\ell \leq [\mathcal{G}_i] \cap \mathcal{G}_j$, all the resulting orbits have such stabilizers. This finishes the proof of the first claim.

(ii) If $[\mathcal{G}_i] \not\leq [\mathcal{G}_j]$ and $[\mathcal{G}_j] \not\leq [\mathcal{G}_i]$, then $[\mathcal{G}_\ell]$ which is a subgroup of both groups is strictly smaller than both, which means one of the resulting orbits must be larger than both input orbits. Next, w.l.o.g., assume $[\mathcal{G}_i] \leq [\mathcal{G}_j]$. Consider proof by contradiction: suppose the product does not have a strictly larger orbit. It follows that $\mathbf{m}_{[\mathcal{G}_j\backslash\mathcal{G}]} \odot \mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]} = \delta^i_{i,i} \mathbf{m}_{[\mathcal{G}_i\backslash\mathcal{G}]}$ for some $\delta^i_{ii} > 0$. Consider the first and i^{th} element of the elementwise product above:

$$|\mathcal{G} : \mathcal{G}_j| \times |\mathcal{G} : \mathcal{G}_i| = \delta^i_{ii} |\mathcal{G} : \mathcal{G}_i|$$
$$\mathbf{m}_{[\mathcal{G}_j\backslash\mathcal{G}]}(i) \times |\mathcal{G} : N_{\mathcal{G}}(\mathcal{G}_i)| = \delta^i_{ii} |\mathcal{G} : N_{\mathcal{G}}(\mathcal{G}_i)|$$

Substituting $\delta^i_{ii} = |\mathcal{G} : \mathcal{G}_j|$ from the first equation into the second equation and simplifying we get $\mathbf{m}_{[\mathcal{G}_j\backslash\mathcal{G}]}(i) = |\mathcal{G} : \mathcal{G}_j|$. This means the action of $\mathcal{G}_i$ on $[\mathcal{G}_j\backslash\mathcal{G}]$ fixes all points, and therefore $\mathcal{G}_i \subseteq \text{Core}_{\mathcal{G}}(\mathcal{G}_j)$ as defined in (21). This contradicts the assumption of (ii). □

A sufficient condition for (ii) in Theorem 5.1 is for the $\mathcal{G}$-action on input $\mathcal{G}$-sets to be faithful. Note that in this case the the core is trivial; see Section 4.1. An implication of this theorem is that repeated self-product $[\mathcal{H}\backslash\mathcal{G}]^D$ is bound to produce a regular orbit. This leads to Theorem 3.3, that we saw earlier. Here, we give a shorter proof using Theorem 5.1; see Fig. 2.

Alternative Proof of Theorem 3.3. Since $\mathcal{G}$ acts faithfully on $\mathbb{N}$, $\text{Core}_{\mathcal{G}}(\mathcal{H}) = \{e\}$. From Theorem 5.1 it follows that each time we calculate a product by $\mathbb{N}$, a strictly smaller stabilizer is produced so that $\mathcal{H} = \mathcal{H}^{(t=0)} > \mathcal{H}^{(1)} > \ldots > \mathcal{H}^{(D)} = \{e\}$, where $\mathcal{H}^{(d)}$ is the smallest stabilizer at time-step d. From Lagrange theorem, the size of a proper subgroup is at most half the size of its overgroup in this sequence of stabilizers. It follows that for any $D \geq \log_2 |\mathcal{H}|$, $[\mathcal{H}\backslash\mathcal{G}]^D$ has an orbit with $\mathcal{H}^{t=D} = \{e\}$ as its stabilizer. □

Example 3 (UNIVERSAL APPROXIMATION FOR $\mathcal{A}_5$)**.** *The alternating group $\mathcal{A}_5$ is the group of even permutations of 5 objects. One way to create a universal approximator for this group to have a regular layer (see Theorem 3.2). A more convenient alternative is to consider the canonical action of this group on a set $\mathbb{N}$ of size 5, and use an order D layer to ensure universality. Using Corollary 3 we get $D \geq 5 = \lceil (3\frac{1}{2}\log_2(4)) - 4\log_2(e) \rceil$. The natural action of $\mathcal{A}_5$ on $\mathbb{N} = [5]$ is isomorphic to $[\mathcal{A}_4\backslash\mathcal{A}_5]$ – i.e., $\mathcal{A}_4$ is a stabilizer. Using this stabilizer in Theorem 3.3, we get the same bound $D \geq 5 = \lceil \log_2(|\mathcal{A}_4|) \rceil$.*

However, using the table of marks we can show that $D = 3$ already produces a regular orbit in this case. The table of marks for the alternating group $\mathcal{A}_5$ is shown in Table 2. Our objective is to find the decomposition of $[\mathcal{A}_4\backslash\mathcal{A}_5]^3$. We do this in steps, first showing

$$[\mathcal{A}_4\backslash\mathcal{A}_5]^2 = [\mathcal{A}_4\backslash\mathcal{A}_5] \cup [\mathcal{C}_3\backslash\mathcal{A}_5] \tag{30}$$

To see this, note that the element-wise product of the vector of marks $\mathbf{m}_{[\mathcal{A}_4\backslash\mathcal{A}_5]}$ (which is next to last row in Table 2) with itself is equal to $\mathbf{m}_{[\mathcal{A}_4\backslash\mathcal{A}_5]} + \mathbf{m}_{[\mathcal{C}_3\backslash\mathcal{A}_5]}$. Since the vector of marks is an injective homomorphism, this implies (30). Applying the same idea one more time, gives

$$\begin{aligned}[\mathcal{A}_4\backslash\mathcal{A}_5]^3 &= ([\mathcal{A}_4\backslash\mathcal{A}_5] \cup [\mathcal{C}_3\backslash\mathcal{A}_5]) \times [\mathcal{A}_4\backslash\mathcal{A}_5] \\ &= 2[\mathcal{A}_4\backslash\mathcal{A}_5] \cup [\mathcal{C}_3\backslash\mathcal{A}_5] \cup [\{e\}\backslash\mathcal{A}_5].\end{aligned}$$

This shows that $[\mathcal{A}_4\backslash\mathcal{A}_5]^3$ contains a regular orbit $[\{e\}\backslash\mathcal{A}_5]$. Therefore, using an order $D = 3$ hidden layer $\mathbb{N}^3$ on which $\mathcal{A}_5$ acts using even permutations, also produces a universal equivariant (invariant) approximator.

Acknowledgements

We thank anonymous reviewers for their constructive feedback. In particular the first proof for Theorem 3.3, as well as clarifications on the proof of the main theorems was proposed by reviewers. This research is in part funded by the Canada CIFAR AI Chair Program.

References

Albooyeh, M., Bertolini, D., and Ravanbakhsh, S. Incidence networks for geometric deep learning. *arXiv preprint arXiv:1905.11460*, 2019.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Burnside, W. *Theory of groups of finite order*. University, 1911.

Cohen, T. S. and Welling, M. Group equivariant convolutional networks. *arXiv preprint arXiv:1602.07576*, 2016a.

Cohen, T. S. and Welling, M. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant cnns on homogeneous spaces. In *Advances in Neural Information Processing Systems*, pp. 9142–9153, 2019a.

Cohen, T. S., Weiler, M., Kicanaoglu, B., and Welling, M. Gauge equivariant convolutional networks and the icosahedral cnn. *arXiv preprint arXiv:1902.04615*, 2019b.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Dieck, T. T. *Transformation groups and representation theory*, volume 766. Springer, 2006.

Dieleman, S., De Fauw, J., and Kavukcuoglu, K. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016.

Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3): 183–192, 1989.

Gens, R. and Domingos, P. M. Deep symmetry networks. In *Advances in neural information processing systems*, pp. 2537–2545, 2014.

Graham, D. and Ravanbakhsh, S. Deep models for relational databases. *arXiv preprint arXiv:1903.09033*, 2019.

Hartford, J., Graham, D. R., Leyton-Brown, K., and Ravanbakhsh, S. Deep models of interactions across sets. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1909–1918, 2018.

Hinton, G. E., Krizhevsky, A., and Wang, S. D. Transforming auto-encoders. In *International conference on artificial neural networks*, pp. 44–51. Springer, 2011.

Hinton, G. E., Sabour, S., and Frosst, N. Matrix capsules with em routing. 2018.

Hornik, K., Stinchcombe, M., White, H., et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. In *Advances in neural information processing systems*, pp. 2017–2025, 2015.

Keriven, N. and Peyré, G. Universal invariant and equivariant graph neural networks. In *Advances in Neural Information Processing Systems*, pp. 7090–7099, 2019.

Kondor, R. and Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.

Kondor, R., Son, H. T., Pan, H., Anderson, B., and Trivedi, S. Covariant compositional networks for learning graphs. *arXiv preprint arXiv:1801.02144*, 2018.

Lenssen, J. E., Fey, M., and Libuschewski, P. Group equivariant capsule networks. *arXiv preprint arXiv:1806.05086*, 2018.

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

Maron, H., Fetaya, E., Segol, N., and Lipman, Y. On the universality of invariant networks. *arXiv preprint arXiv:1901.09342*, 2019.

Maron, H., Litany, O., Chechik, G., and Fetaya, E. On learning sets of symmetric elements. *arXiv preprint arXiv:2002.08599*, 2020.

Minsky, M. and Papert, S. A. *Perceptrons: An introduction to computational geometry*. MIT press, 2017.

Pfeiffer, G. The subgroups of m24, or how to compute the table of marks of a finite group. *Experimental Mathematics*, 6(3):247–270, 1997.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Ravanbakhsh, S., Schneider, J., and Poczos, B. Deep learning with sets and point clouds. In *International Conference on Learning Representations (ICLR) – workshop track*, 2017a.

Ravanbakhsh, S., Schneider, J., and Poczos, B. Equivariance through parameter-sharing. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *JMLR: WCP*, August 2017b.

Rotman, J. J. *An introduction to the theory of groups*, volume 148. Springer Science & Business Media, 2012.

Sabour, S., Frosst, N., and Hinton, G. E. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pp. 3856–3866, 2017.

Sannai, A., Takai, Y., and Cordonnier, M. Universal approximations of permutation invariant/equivariant functions by deep neural networks. *arXiv preprint arXiv:1903.01939*, 2019.

Segol, N. and Lipman, Y. On universal equivariant set networks. *arXiv preprint arXiv:1910.02421*, 2019.

Shawe-Taylor, J. Building symmetries into feedforward networks. In *Artificial Neural Networks, 1989., First IEE International Conference on (Conf. Publ. No. 313)*, pp. 158–162. IET, 1989.

Shawe-Taylor, J. Symmetries and discriminability in feedforward network architectures. *IEEE Transactions on Neural Networks*, 4(5):816–826, 1993.

Sturmfels, B. *Algorithms in invariant theory*. Springer Science & Business Media, 2008.

Weiler, M. and Cesa, G. General e (2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, pp. 14334–14345, 2019.

Wood, J. Invariant pattern recognition: a review. *Pattern recognition*, 29(1):1–17, 1996.

Wood, J. and Shawe-Taylor, J. Representation theory and invariant neural networks. *Discrete applied mathematics*, 69(1-2):33–60, 1996.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., and Brostow, G. J. Harmonic networks: Deep translation and rotation equivariance. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

Yarotsky, D. Universal approximations of invariant maps by neural networks. *arXiv preprint arXiv:1804.10306*, 2018.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.