On the Universality of Invariant Networks

Haggai Maron [1] **Ethan Fetaya** [2 3] **Nimrod Segol** [1] **Yaron Lipman** [1]

Abstract

Constraining linear layers in neural networks to respect symmetry transformations from a group G is a common design principle for invariant networks that has found many applications in machine learning. In this paper, we consider a fundamental question that has received little attention to date: Can these networks approximate any (continuous) invariant function? We tackle the rather general case where $G \leq S_n$ (an arbitrary subgroup of the symmetric group) that acts on $\mathbb{R}^n$ by permuting coordinates. This setting includes several recent popular invariant networks. We present two main results: First, G-invariant networks are universal if high-order tensors are allowed. Second, there are groups G for which higher-order tensors are unavoidable for obtaining universality. G-invariant networks consisting of only first-order tensors are of special interest due to their practical value. We conclude the paper by proving a necessary condition for the universality of G-invariant networks that incorporate only first-order tensors.

1. Introduction

The basic paradigm of deep neural networks is repeatedly composing ”layers” of linear functions with non-linear, entrywise activation functions to create effective predictive models for learning tasks of interest.

When trying to learn a function (task) f that is known to be invariant to some group of symmetries G (i.e., G-invariant function) it is common to use linear layers that respect this symmetry, namely, invariant and/or equivariant linear layers. Networks with invariant/equivariant linear layers with respect to some group G will be referred here as G-*invariant networks*.

A fundamental question in learning theory is that of *approximation* or *universality* (Cybenko, 1989; Hornik, 1991). In the invariant case: Can a G-invariant network approximate an arbitrary continuous G-invariant function?

The goal of this paper is to address this question for *all* finite permutation groups $G \leq S_n$, where S_n is the symmetric group acting on $[n] = \{1, 2, \dots, n\}$. Note that this is a fairly general setting that contains many useful examples (detailed below).

The archetypal example of G-invariant networks is Convolutional Neural Networks (CNNs) (LeCun et al., 1989; Krizhevsky et al., 2012) that restrict their linear layers to convolutions in order to learn image tasks that are translation invariant or equivariant [1].

In recent years researchers are considering other types of data and/or symmetries and consequently new G-invariant networks have emerged. Tasks involving point clouds or sets are in general invariant to the order of the input and therefore permutation invariance/equivariance was developed (Qi et al., 2017; Zaheer et al., 2017). Learning tasks involving interaction between different sets, where the input data is tabular, require dealing with different permutations acting independently on each set (Hartford et al., 2018). Tasks involving graphs and hyper-graphs lead to symmetries defined by tensor products of permutations (Kondor et al., 2018; Maron et al., 2019). A general treatment of invariance/equivariance to finite subgroups of the symmetric group is discussed in (Ravanbakhsh et al., 2017); infinite symmetries are discussed in general in (Kondor & Trivedi, 2018) as well as in (Cohen & Welling, 2016a;b; Cohen et al., 2018; Weiler et al., 2018).

Among these examples, universality is known for point-clouds networks and sets networks (Qi et al., 2017; Zaheer et al., 2017), as well as networks invariant to finite translation groups (e.g., cyclic convolutional neural networks) (Yarotsky, 2018). However, universality is not known for tabular and multi-set networks (Hartford et al., 2018), graph

[1] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel [2] Department of Computer Science, University of Toronto, Toronto, Canada [3] Vector Institute. Correspondence to: Haggai Maron <haggai.maron@weizmann.ac.il>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] It is common to use convolutional layers without cyclic padding which implies that these networks are not precisely translation invariant.

and hyper-graph networks (Kondor et al., 2018; Maron et al., 2019); and networks invariant to finite translations with rotations and/or reflections. We cover all these cases in this paper.

Maybe the most related work to ours is (Yarotsky, 2018) that considered actions of compact groups and suggested provably universal architectures that are based on polynomial layers. In contrast, we study the standard and widely used linear layer model.

The paper is organized as follows: First, we prove that an arbitrary continuous function $f : \mathbb{R}^n \to \mathbb{R}$ invariant to an arbitrary permutation group $G \leq S_n$ can be approximated using a G-invariant network. The proof is constructive and makes use of linear equivariant layers between tensors $\mathbf{X} \in \mathbb{R}^{n^k}$ of order $k \leq d$, where d depends on the permutation group G.

Second, we prove a lower bound on the order d of tensors used in a G-invariant network so to achieve universality. Specifically, we show that for $G = A_n$ (the alternating group) any G-invariant network that uses tensors of order at-most $d = (n-2)/2$ cannot approximate arbitrary G-invariant functions.

We conclude the paper by considering the question: For which groups $G \leq S_n$, G-invariant networks using only first order tensors are universal? We prove a necessary condition, and describe families of groups for which universality cannot be attained using only first order tensors.

2. Preliminaries and main results

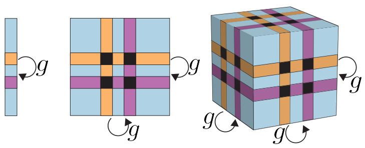
The symmetries we consider in this paper are arbitrary subgroups of the symmetric group, i.e., $G \leq S_n$. The action of G on $x \in \mathbb{R}^n$ used in this paper is defined as

$$g \cdot x = (x_{g^{-1}(1)}, \dots, x_{g^{-1}(n)}), \; g \in G. \tag{1}$$

The action of G on *tensors* $\mathbf{X} \in \mathbb{R}^{n^k \times a}$ (the last index, denoted j represents feature depth) is defined similarly by

$$(g \cdot \mathbf{X})_{i_1 \dots i_k, j} = \mathbf{X}_{g^{-1}(i_1) \dots g^{-1}(i_k), j}, \; g \in G. \tag{2}$$

The inset illustrates this action on tensors of order $k = 1, 2, 3$: the permutation g is a transposition of two numbers and is applied to each dimension of the tensor.



Definition 1. *A G-invariant function is a function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies $f(g \cdot x) = f(x)$ for all $x \in \mathbb{R}^n$ and $g \in G$.*

Definition 2. *A linear equivariant layer is an affine map $L : \mathbb{R}^{n^k \times a} \to \mathbb{R}^{n^l \times b}$ satisfying $L(g \cdot \mathbf{X}) = g \cdot L(\mathbf{X})$, for all $g \in G$, and $\mathbf{X} \in \mathbb{R}^{n^k \times a}$. An invariant linear layer is an affine map $h : \mathbb{R}^{n^k \times a} \to \mathbb{R}^b$ satisfying $h(g \cdot \mathbf{X}) = h(\mathbf{X})$, for all $g \in G$, and $\mathbf{X} \in \mathbb{R}^{n^k \times a}$.*

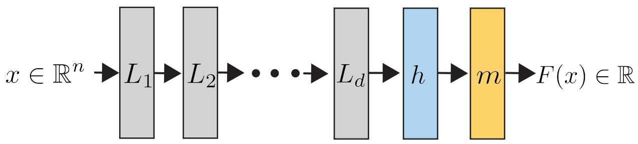


Figure 1. Illustration of invariant network architecture. The function is composed of multiple linear G-equivariant layers (gray), possibly of high order, and ends with a linear G-invariant function (light blue) followed by a Multi Layer Perceptron (yellow).

A common way to construct G-invariant networks is:

Definition 3. *A G-invariant network is a function $F : \mathbb{R}^{n \times a} \to \mathbb{R}$ defined as*

$$F = m \circ h \circ L_d \circ \sigma \circ \dots \circ \sigma \circ L_1,$$

where L_i are linear G-equivariant layers, σ is an activation function [2]*, h is a G-invariant layer, and m is a Multi-Layer Perceptron (MLP).*

Figure 1 illustrates the G-invariant network model. By construction, G-invariant networks are G-invariant functions (note that entrywise activation is equivariant as-well). This framework has been used, with appropriate group G, in previous works to build predictive G-invariant models for learning.

Our goal is to show the *approximation power* of G-invariant networks. Namely, that G-invariant networks can approximate arbitrary continuous G-invariant functions f. Without loss of generality, we consider only functions of the form $f : \mathbb{R}^n \to \mathbb{R}$. Indeed, in case of multiple features, $\mathbb{R}^{n \times a}$, we rearrange the input as $\mathbb{R}^{n'}$, $n' = na$, and take the appropriate $G' \leq S_{n'}$. We prove:

Theorem 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous G-invariant function for some $G \leq S_n$, and $K \subset \mathbb{R}^n$ a compact set. There exists a G-invariant network that approximates f to an arbitrary precision.*

The proof of Theorem 1 is constructive and builds an f-approximating G-invariant network with hidden tensors $\mathbf{X} \in \mathbb{R}^{n^d}$ of order d, where $d = d(G)$ is a natural number depending on the group G. Unfortunately, we show that in the worst case d can be as high as $\frac{n(n-1)}{2}$. Note that $d = 2$ could already be computationally challenging. It is therefore of interest to ask whether there exist more efficient

[2] We assume any activation function for which the universal approximation theorem for MLP holds, e.g., ReLU and sigmoid.

G-invariant networks that use lower order tensors without sacrificing approximation power. Surprisingly, the answer is that in general we can not go lower than order n for general permutation groups G. Specifically, we prove the following for $G = A_n$, the alternating group:

Theorem 2. *If an A_n-invariant network has the universal approximation property then it consists of tensors of order at least $\frac{n-2}{2}$.*

Although in general we cannot expect universal approximation of G-invariant networks with inner tensor order smaller than $\frac{n-2}{2}$, it is still possible that for *specific* groups of interest we can prove approximation power with more efficient (i.e., lower order inner tensors) G-invariant networks. Of specific interest are G-invariant networks that use only first order tensors. In section 5 we prove the following necessary condition for universality of first-order G-invariant networks:

Theorem 3. *Let $G \leq S_n$. If first order G-invariant networks are universal, then $\left|[n]^2/H\right| < \left|[n]^2/G\right|$ for any strict super-group $G < H \leq S_n$.*

$|[n]^2/G|$ is the number of equivalence classes of $[n]^2$ defined by the relation: $(i_1, i_2) \sim (j_1, j_2)$ if $j_\ell = g(i_\ell)$, $\ell = 1, 2$ for some $g \in G$. Intuitively, this condition asks that super-groups of G have *strictly* better separation of the double index space $[n]^2$.

3. G-invariant networks universality

The key to showing theorem 1, namely that G-invariant networks are universal, is showing they can approximate a set of functions that are: (i) G-invariant; and (ii) can approximate arbitrary G-invariant functions to a desired precision. The G-invariant polynomials are an example of such a set:

Definition 4. *The G-invariant polynomials are all the polynomials in $x_1, \dots, x_n$ over $\mathbb{R}$ that are also G-invariant functions. They are denoted $\mathbb{R}[x_1, \dots, x_n]^G$, where $\mathbb{R}[x_1, \dots, x_n]$ is the set of all polynomials over $\mathbb{R}$.*

To see that G-invariant polynomials can approximate any arbitrary (continuous) function $f : K \subset \mathbb{R}^n \to \mathbb{R}$, where K is a compact set, one can use the Stone-Weiestrass (SW) theorem, as done in (Yarotsky, 2018): First use SW to approximate f over a symmetrized domain $K' = \cup_{g \in G} g \cdot K$ by *some* (not necessarily G-invariant) polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$. Second, consider

$$q(x) = \frac{1}{|G|} \sum_{g \in G} p(g \cdot x).$$

q is a G-invariant polynomial and hence

$$q \in \mathbb{R}[x_1, \dots, x_n]^G,$$

furthermore for $x \in K$:

$$|q(x) - f(x)| \leq$$
$$\frac{1}{|G|} \sum_{g \in G} \left|p(g \cdot x) - f(g \cdot x)\right| \leq \max_{x \in K'} |p(x) - f(x)|.$$

Our goal in this section is to prove the following proposition that, together with the comment above, prove theorem 1:

Proposition 1. *For any $\epsilon > 0$, $K \subset \mathbb{R}^n$ compact set, and G-invariant polynomial $p \in \mathbb{R}[x_1, \dots, x_n]^G$ there exists a G-invariant network F that approximates p to an ϵ-accuracy, namely $\max_{x \in K} |F(x) - p(x)| < \epsilon$.*

The proposition will be proved in several steps:

(i) We represent p as $p(x) = \sum_{k=0}^{d} p_k(x)$, where p_k is a G-invariant *homogeneous* polynomial of degree k, i.e., $p_k \in \mathbb{R}_k[x_1, \dots, x_n]^G$.

(ii) We characterize all homogeneous G-invariant polynomials of a fixed degree k. In particular we find a basis to all such polynomials, $b_{k1}, b_{k2}, \dots, b_{kn_k} \in \mathbb{R}_k[x_1, \dots, x_n]^G$. Using the bases of homogeneous G-invariant polynomials of degrees up-to d we write

$$p(x) = \sum_{k=0}^{d} \sum_{j=1}^{n_k} \alpha_{kj} b_{kj}(x). \tag{3}$$

(iii) We approximate each basis element b_{kj} using a G-invariant network.

(iv) We construct a G-invariant network F approximating p to an ϵ-accuracy using Equation 3 and (iii).

3.1. Proof of proposition 1

Part (i): It is a known fact that a G-invariant polynomial can be written as a sum of homogeneous G-invariant polynomials (Kraft & Procesi, 2000):

Lemma 1. *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a G-invariant polynomial of degree d. Then p can be written as $p(x) = \sum_{k=0}^{d} p_k(x)$ where p_k are homogeneous G-invariant polynomials of degree k.*

Part (ii): We need to find bases for the linear spaces of homogeneous G-invariant polynomials of degree $k = 0, 1, \dots, d$, i.e., $\mathbb{R}_k[x_1, \dots, x_n]^G$. Any homogeneous polynomial of degree k can be written as

$$p(x) = \sum_{i_1, \dots, i_k = 1}^{n} \mathbf{W}_{i_1 \dots i_k} x_{i_1} \cdots x_{i_k}, \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{n^k}$ is its coefficient tensor; since $x_{i_1} \cdots x_{i_k} = x_{i_{\sigma(1)}} \cdots x_{i_{\sigma(k)}}$ for all $\sigma \in S_k$, a unique choice of $\mathbf{W}$ can be obtained by taking a symmetric $\mathbf{W}$. That is, $\mathbf{W}$ that satisfies $\mathbf{W}_{i_1\cdots i_k} = \mathbf{W}_{i_{\sigma(1)}\cdots i_{\sigma(k)}}$, for all $\sigma \in S_k$. In short, we ask $\mathbf{W} \in \mathrm{Sym}^k(\mathbb{R}^n) \subset \mathbb{R}^{n^k}$. For example, the case $k = 2$ amounts to representing a quadratic form using a symmetric matrix, that is $\mathbf{W}$ satisfies in this case $\mathbf{W} = \mathbf{W}^T$. The next proposition shows that if p is G-invariant, its coefficient tensor is a fixed point of the action of G on symmetric tensors $\mathbf{W} \in \mathbb{R}^{n^k}$:

Proposition 2. *Let $p \in \mathbb{R}_k[x_1, \ldots, x_n]^G$. Then its coefficient tensor $\mathbf{W} \in \mathbb{R}^{n^k}$ satisfies the fixed point equation:*

$$g \cdot \mathbf{W} = \mathbf{W}, \ \forall g \in G. \tag{5}$$

Proof. From the fact that p is G-invariant we get the following set of equations $p(x) = p(g \cdot x)$, for all $g \in G$.

$$\begin{aligned} p(x) &= p(g \cdot x) \\ &= \sum_{i_1,\ldots,i_k=1}^{n} \mathbf{W}_{i_1\ldots i_k}\, x_{g^{-1}(i_1)} \cdots x_{g^{-1}(i_k)} \\ &= \sum_{i_1,\ldots,i_k=1}^{n} \mathbf{W}_{g(i_1)\ldots g(i_k)}\, x_{i_1} \cdots x_{i_k}. \end{aligned}$$

By equating monomials' coefficients of $p(x)$ and $p(g \cdot x)$ and the symmetry of $\mathbf{W}$ we get

$$\mathbf{W}_{i_1\ldots i_k} = \mathbf{W}_{g(i_1)\ldots g(i_k)}.$$

This implies that $\mathbf{W}$ satisfies $g \cdot \mathbf{W} = \mathbf{W}$ for all $g \in G$. □

Equation 5 is a linear homogeneous system of equations and therefore the set of solutions $\mathbf{W}$ forms a linear space. To define a basis for this linear space we first define the following equivalence relation: $(i_1, \ldots, i_k) \sim (j_1, \ldots, j_k)$ if there exists $g \in G$ and $\sigma \in S_k$ so that $j_\ell = g(i_{\sigma(\ell)})$, $\ell = 1, \ldots, k$. Intuitively, g takes care of the G-invariance while σ factors out the fact that the monomials $x_{i_1} \cdots x_{i_k} = x_{\sigma(i_1)} \cdots x_{\sigma(i_k)}$. For example, let $n = 5$, $k = 3$, $g = (23)(45)$, $\sigma = (23)$ (we use cycle notation), then we have: $(2, 2, 4) \sim (3, 5, 3)$. The equivalence classes are denoted τ and called the *k-classes*. We show:

Proposition 3. *The set of polynomials*

$$p^\tau(x) = \sum_{(i_1,\ldots,i_k)\in\tau} x_{i_1} \cdots x_{i_k}, \tag{6}$$

where τ is a k-class, form a basis to $\mathbb{R}_k[x_1, \ldots, x_n]^G$.

Proof. Denote $\mathbf{W}^\tau$ the symmetric coefficient tensor of p^τ, Note that

$$\mathbf{W}^\tau_{i_1\ldots i_k} = \begin{cases} 1 & (i_1, \ldots, i_k) \in \tau \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

Since

$$p^\tau(g \cdot x) = \sum_{(i_1,\ldots,i_k)\in\tau} x_{g^{-1}(i_1)} \cdots x_{g^{-1}(i_k)} = p^\tau(x),$$

$p^\tau \in \mathbb{R}_k[x_1, \ldots, x_n]^G$. The set of polynomials p^τ, with τ a k-classes, is a linearly independent set since each p^τ contains a different collection of monomials. By Proposition 2, the symmetric coefficient tensor $\mathbf{W}$ of every $q \in \mathbb{R}_k[x_1, \ldots, x_n]^G$ satisfies the fixed-point equation, equation 5. This in particular means that $\mathbf{W}$ is constant on its k-classes. Hence $\mathbf{W}$ can be written as linear combination of $\mathbf{W}^\tau$, see also equation 7. □

As we later show, the fixed point equation, equation 5, is also used to characterize and compute a basis for the space of *linear* permutation-equivariant and invariant layers (Maron et al., 2019). These equations are equivalently formulated using weight sharing scheme in (Ravanbakhsh et al., 2017). A slight difference in this case, that deals with polynomials, is the additional constraints that formulate the symmetry of $\mathbf{W}$ which are needed since every polynomial of degree > 1 has several representing tensors $\mathbf{W}$.

Part (iii): Our next step is approximating each p^τ with a G-invariant network. The next proposition introduces the building blocks of this construction:

Proposition 4. *Let τ be a k-class and let $L^\tau_\ell : \mathbb{R}^n \to \mathbb{R}^{n^k}$, $\ell = 1, \ldots, k$, be a linear operator defined as follows: For $x \in \mathbb{R}^n$*

$$L^\tau_\ell(x)_{i_1\ldots i_k} = \begin{cases} x_{i_\ell} & (i_1, \ldots, i_k) \in \tau \\ 0 & \textit{otherwise} \end{cases}.$$

Then L^τ_ℓ is a linear G-equivariant function, that is

$$L^\tau_\ell(g \cdot x) = g \cdot L^\tau_\ell(x), \ \forall x \in \mathbb{R}^n, g \in G.$$

Proof. We have :

$$g \cdot L^\tau_\ell(x)_{i_1\ldots i_k} = \begin{cases} x_{g^{-1}(i_\ell)} & (g^{-1}(i_1), \ldots, g^{-1}(i_k)) \in \tau \\ 0 & \text{otherwise} \end{cases}$$

On the other hand,

$$L^\tau_\ell(g \cdot x)_{i_1\ldots i_k} = \begin{cases} x_{g^{-1}(i_\ell)} & (i_1, \ldots, i_k) \in \tau \\ 0 & \text{otherwise} \end{cases}$$

and both expressions are equal since $(i_1, \ldots, i_k) \in \tau$ if and only if $(g^{-1}(i_1), \ldots, g^{-1}(i_k)) \in \tau$ by definition of τ. □

Next, we construct the approximating G-invariant network:

Proposition 5. *For any $\epsilon > 0$, $K \subset \mathbb{R}^n$ compact set, and τ k-class there exists a G-invariant network F^τ that approximates p^τ from equation 6 to an ϵ-accuracy.*

Proof. Let $c > 0$ be sufficiently large so that $K \subset [-c,c]^n \subset \mathbb{R}^n$. Denote $m^k : \mathbb{R}^k \to \mathbb{R}$ an MLP that approximates the multiplication function, $f(y_1,\dots,y_k) = \prod_{i=1}^k y_i$, in $[-c,c]^k$ to $n^{-k}\epsilon$-accuracy, i.e., $\max_{-c\le y_i \le c} \left| f(y) - m^k(y) \right| < n^{-k}\epsilon$.

Consider the following G-invariant network: First, given an input $x \in \mathbb{R}^n$ map it to $\mathbb{R}^{n^k \times k}$ (i.e., k is the number of channnels) by

$$L^\tau(x)_{i_1\dots i_k,\ell} = L^\tau_\ell(x)_{i_1\dots i_k}. \tag{8}$$

$L^\tau : \mathbb{R}^n \to \mathbb{R}^{n^k \times k}$ is a linear equivariant layer (see equation 2). Second, apply m^k to the feature dimension in $\mathbb{R}^{n^k\times k}$. That is, given $y \in \mathbb{R}^{n^k \times k}$ define

$$M^k(y)_{i_1,\dots,i_k} = m^k(y_{i_1\dots,i_k,1},\dots,y_{i_1\dots,i_k,k}).$$

Note that $M^k : \mathbb{R}^{n^k\times k} \to \mathbb{R}^{n^k}$ can be interpreted as a composition of equivariant linear layers [3].

Lastly, denote $s : \mathbb{R}^{n^k} \to \mathbb{R}$ the summation layer: for $z \in \mathbb{R}^{n^k}$, $s(z) = \sum_{i_1\dots i_k=1}^n z_{i_1\dots i_k}$. Note that M^k, s are equivariant, invariant (respectively) for all $G \le S_n$. This construction can be visualized using the following diagram:

$$\begin{tikzcd} \mathbb{R}^n \arrow[r, "L^\tau"] & \mathbb{R}^{n^k\times k} \arrow[r, "M^k"] & \mathbb{R}^{n^k} \arrow[r, "s"] & \mathbb{R} \end{tikzcd}$$

This G-invariant network $F^\tau = s \circ M^k \circ L^\tau$ approximates p^τ to an ϵ-accuracy over the compact set $K \subset \mathbb{R}^n$. Indeed, let $x \in K$, then

$$\begin{aligned}
&\left|F^\tau(x) - p^\tau(x)\right| \\
&\le \sum_{i_1\dots i_k=1}^{n} \left| M^k(L^\tau(x))_{i_1\dots i_k} - \mathbf{W}^\tau_{i_1\dots i_k} x_{i_1}\cdots x_{i_k}\right| \\
&\le \sum_{i_1\dots i_k=1}^{n} \begin{cases} \left|m^k(x_{i_1},\dots,x_{i_k}) - x_{i_1}\cdots x_{i_k}\right| & {\scriptstyle (i_1,\dots,i_k)\in\tau} \\ 0 & \text{otherwise} \end{cases} \\
&\le \epsilon,
\end{aligned}$$

where in the last inequality we used the $n^{-k}\epsilon$-accuracy of m^k to the product operator in $[-c,c]^k \subset \mathbb{R}^k$.

□

Part (iv): In the final stage, we would like to approximate an arbitrary $p \in \mathbb{R}[x_1,\dots,x_n]^G$ with a G-invariant network to ϵ-accuracy over a compact set $K \subset \mathbb{R}^n$.

Proof. (proposition 1) Let us denote by $b_{k1},\dots,b_{kn_k}$ the polynomials p^τ, with τ the k-classes. Let F^{kj} denote the G-invariant network approximating b_{kj}, $k = 0,1,\dots d$, $j \in [n_k]$, to an ϵ-accuracy over the set K, the existence of which is guaranteed by proposition 5. We now utilize the decomposition of p shown in equation 3 and get

$$\begin{aligned}
&\left| p(x) - \sum_{k=0}^{d}\sum_{j=1}^{n_k} \alpha_{kj} F^{kj}(x) \right| \\
&\le \sum_{k=0}^{d}\sum_{j=1}^{n_k} \left|\alpha_{kj}\right| \left| b_{kj}(x) - F^{kj}(x)\right| \\
&\le \epsilon \left\|\alpha\right\|_1 ,
\end{aligned}$$

where $\|\alpha\|_1 = \sum_{k,j} |\alpha_{kj}|$ depends only upon p, where ϵ is arbitrary. To finish the proof we need to show that $F = \sum_{k=0}^d \sum_{j=1}^{n_k} \alpha_{kj} F^{kj}$ can indeed be realized as a *single, unified* G-invariant network. This is a simple yet technical construction and we defer the proof of this fact to the supplementary material:

Lemma 2. *There exists a G-invariant network in the sense of definition 3 that realizes the sum of G-invariant networks $F = \sum_{k=0}^d \sum_{j=1}^{n_k} \alpha_{kj} F^{kj}$.*

□

3.2. Bounded order construction

We have constructed a G-invariant network F that approximates an arbitrary G-invariant polynomial $p \in \mathbb{R}[x_1,\dots,x_n]^G$ of degree d. The network F uses d-dimensional tensors, where d matches the degree of p. In this subsection we construct a G-invariant network F that approximates p with maximal tensor order that depends only on the group $G \le S_n$. Therefore, the tensor order is independent of the degree of the polynomial p. We use the following theorem by Noether (Kraft & Procesi, 2000):

Theorem 4. ***(Noether)*** *Let G be a finite group acting linearly on $\mathbb{R}^n$. There exist finitely many G-invariant polynomials $f_1, ..., f_m \in \mathbb{R}[x_1,\dots,x_n]^G$ such that any invariant polynomial $p \in \mathbb{R}[x_1,\dots,x_n]^G$ can be expressed as*

$$p(x) = h(f_1(x), ..., f_m(x)),$$

where $h \in \mathbb{R}[x_1,\dots,x_m]$ is a polynomial and $\deg(f_i) \le |G|$, $i = 1,\dots,m$.

The idea of using a set of generating invariant polynomials in the context of universality was introduced in (Yarotsky, 2018).

For the case of interest in this paper, namely $G \le S_n$, there exists a generating set of G-invariant polynomials of degree bounded by $\frac{n(n-1)}{2}$, for $n \ge 3$, see (Göbel, 1995). We can

[3] In fact, any application of an MLP to the feature dimension is G-equivariant for any $G \le S_n$ since it can be realized by scaling of the identity operator, possibly with a constant and non-linear point-wise activations (see e.g.(Qi et al., 2017; Zaheer et al., 2017)).

now repeat the construction above, building a G-invariant network F_i approximating f_i to a ϵ_1-accuracy, $i = 1, \ldots, m$. The maximal order of these networks is bounded by $d \leq \frac{n(n-1)}{2}$. These networks can be combined, as above, to a single G-invariant network $F : \mathbb{R}^n \to \mathbb{R}^m$ with the final output approximating $f(x) = (f_1(x), \ldots, f_m(x))$ to a ϵ_1-accuracy. Now we compose the output of F with an MLP $H : \mathbb{R}^m \to \mathbb{R}$ approximating the polynomial h over the compact set $f(K) + B_\epsilon \subset \mathbb{R}^m$ to an ϵ-accuracy, where B_ϵ is a closed ball centered at the origin of radius ϵ and the sum is the Minkowski sum. Since H is continuous and $f(K) + B_\epsilon$ is compact, there exists $\delta > 0$ so that $|H(y) - H(y')| \leq \epsilon$ if $\|y - y'\|_2 \leq \delta$. We use $\epsilon_1 = \min\{\delta, \epsilon\}$ for the construction of F above. We have:

$$\begin{aligned} |H(F(x)) - h(f(x))| &\leq |H(F(x)) - H(f(x))| \\ &\quad + |H(f(x)) - h(f(x))| \leq 2\epsilon, \end{aligned}$$

for all $x \in K$. We have constructed $H \circ F$ that is a G-invariant network with maximal tensor order bounded by $\frac{n(n-1)}{2}$ approximating p to an arbitrary precision.

3.3. Examples

Universality of (hyper-) graph networks. Graph, or hyper-graph data can be described using tensors $\mathbf{X} \in \mathbb{R}^{n^k \times a}$, where n is the number of vertices of the graph and $x_{i_1,i_2,\ldots,i_k,:} \in \mathbb{R}^a$ is a feature vector attached to a (generalized-)edge defined by the ordered set of vertices $(i_1, i_2, \ldots, i_k)$. For example, an adjacency matrix of an n-vertex graph is described by $\boldsymbol{X} \in \mathbb{R}^{n^2}$. The graph symmetries are reordering the vertices by a permutation, namely $g \cdot \mathbf{X}$, where $g \in S_n$. Typically, any function we would like to learn on graphs would be invariant to this action. Recently, (Maron et al., 2019) characterized the spaces of equivariant and invariant linear layers with this symmetry, provided a formula for their basis and employed the corresponding G-invariant networks for learning graph-related tasks. A corollary of Theorem 1 is that this construction yields a universal approximator of continuous functions defined on graphs. This is in contrast to the popular *message passing neural network* model (Gilmer et al., 2017) that was recently shown to be non-universal (Xu et al., 2019).

Universality of rotation invariant convolutional networks. For learning tasks involving $m \times m$ images one might require invariance to periodic translations and 90 degree rotations. Note that periodic translations and 90 degree rotations can be seen as permutations in S_n, $n = m^2$, acting on the pixels of the image. Constructing a suitable G-invariant network would lead, according to Theorem 1, to a universal approximator.

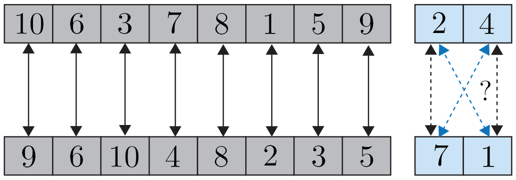


Figure 2. Illustration of the $(n-2)$-transitivity of A_n, the main property we use in this section. Any subset of distinct $n - 2$ elements can be mapped to any other subset of distinct $n - 2$ elements (gray). If needed, a transposition can be applied to the remaining 2 elements (blue) to assure an even permutation.

4. A lower bound on equivariant layer order

In the previous section we showed how an arbitrary G-invariant polynomial can be approximated with a G-invariant network with tensor order $d = d(G) \leq \frac{n(n-1)}{2}$. This upper-bound would be prohibitive in practice. In this section we prove a *lower bound*: We show that there exists a group for which the tensor order cannot be less than $\frac{n-2}{2}$ if we wish to maintain the universal approximation property.

We consider the alternating group, $G = A_n \leq S_n$. Remember that $g \in A_n$ if g has an even number of transpositions.

Definition 5. *A group $G \leq S_n$ is k-transitive if for every two sequences $(i_1, i_2, \ldots, i_k)$, $(j_1, j_2, \ldots, j_k)$ of distinct elements in $[n]$ there exists $g \in G$ so that $j_\ell = g(i_\ell)$, for $\ell = 1, \ldots, k$.*

The alternating group is $(n-2)$-transitive (see figure 2 and (Dixon & Mortimer, 1996)). Our goal is to prove:

Theorem 2. *If an A_n-invariant network has the universal approximation property, then it consists of tensors of order at least $\frac{n-2}{2}$.*

For the proof we first need a characterization of the linear equivariant layers $L : \mathbb{R}^{n^k \times a} \to \mathbb{R}^{n^l \times b}$, where $l = 0$ represents the invariant case. By definition $L(g \cdot \mathbf{X}) = g \cdot L(\mathbf{X})$ for all $\mathbf{X} \in \mathbb{R}^{n^k \times a}$. In particular this means that

$$g^{-1} \cdot L(g \cdot \mathbf{X}) = L(\mathbf{X})$$

Recall that L is an *affine map* (see definition 2) and therefore can be represented as a sum of a purely linear part and a constant part. Representing the linear part of L as a tensor $\mathbf{L} \in \mathbb{R}^{n^{k+l} \times a \times b}$ these equations become the fixed-point equation for linear equivariant layers (see supplementary material for derivation):

$$g \cdot \mathbf{L} = \mathbf{L}, \; g \in G. \tag{9}$$

The constant part of L can be encoded using a tensor $\mathbf{B} \in \mathbb{R}^{n^l \times b}$ that satisfies equation 9 as-well. Note the that this fixed point equation is similar to the fixed point equation of homogeneous G-invariant polynomials, equation 5. We denote by $\mathcal{L}^G$ the collection of $L : \mathbb{R}^{n^k \times a} \to \mathbb{R}^{n^l \times b}$ linear G-equivariant ($l > 0$) or G-invariant ($l = 0$) layers.

Proposition 6. *If $k + l \leq n - 2$, then $\mathcal{L}^{A_n} = \mathcal{L}^{S_n}$.*

Proof. In view of the fixed point equation for equivariant/invariant layers (9) we need to show the solution set to this equation is identical for $G = A_n$ and $G = S_n$, as long as $k + l \leq n - 2$. The solution set of the fixed point equation consists of tensors $\mathbf{L}$ that are constant on each equivalence class defined by the equivalence relation: $(i_1, \ldots, i_{k+l}) \sim (j_1, \ldots, j_{k+l})$ if $j_\ell = g(i_\ell)$ for $\ell = 1, \ldots, k + l$.

Both A_n and S_n are $(n-2)$-transitive[4]. Therefore, the equivalence relations defined above for A_n and S_n reduce to the *same* equivalence relation $(i_1, \ldots, i_{k+l}) \sim (j_1, \ldots, j_{k+l})$ if $i_\alpha = i_\beta$ if and only if $j_\alpha = j_\beta$, for all $\alpha, \beta \in [k + l]$ (see (Maron et al., 2019) where these classes are called *equality patterns*). Since this equivalence relation is the same for A_n, S_n, we get that the solution set of the fixed point equation (9) is the same for both groups. Since the constant part tensor $\mathbf{B}$ is of smaller order than $k + l \leq n - 2$, the same argumentation applies to the constant part, as-well. □

Proposition 6 implies that any A_n-invariant network with tensor order $\leq (n-2)/2$ will be in fact S_n-invariant. Therefore, one approach to show that such networks have limited approximation power is to come up with an A_n-invariant continuous function that is *not* S_n-invariant, as follows:

Proof. (Theorem 2) Consider the Vandermonde polynomial $V(x) = \prod_{1 \leq i < j \leq n}(x_i - x_j)$. It is not hard to check that V is A_n-invariant but not S_n-invariant (consider, e.g., $g = (12) \in S_n$). Pick $x \in \mathbb{R}^n$ with distinct coordinates. Then it holds that $V(x) \neq 0$. Let $\epsilon > 0$ and $K \subset \mathbb{R}^n$ a compact set containing both $x, g \cdot x$ for $g = (12)$. Assume by way of contradiction that there exists an A_n-invariant network F, which is S_n-invariant due to the above, such that $|V(x) - F(x)| \leq \epsilon$ as well as:

$$\begin{aligned} |V(g \cdot x) - F(g \cdot x)| &= |(-1)V(x) - F(x)| \\ &= |V(x) + F(x)| \\ &\leq \epsilon \end{aligned}$$

These last equations imply that $|V(x)| \leq \epsilon$ and since ϵ is arbitrary we get $V(x) = 0$, a contradiction. □

5. Universality of first order networks

We have seen that G-invariant networks with tensor order $\frac{n(n-1)}{2}$ are universal. On the other hand for general permutation groups G the tensor order is at least $(n - 2)/2$ if universality is required. A particularly important question for applications, where higher order tensors are computationally prohibitive, is which permutation groups G give rise to *first order G-invariant networks* that are universal.

Definition 6. *A first order G-invariant network is a G-invariant network where the maximal tensor order is* 1.

In this section we discuss this (mostly) open question. First, we note that there are a few cases for which first order G-invariant networks are known to be universal: for instance, when $G = \{e\}$ (i.e., the trivial group), G-invariant networks are composed of fully connected layers, a case which is covered by the original universal approximation theorems (Cybenko, 1989; Hornik, 1991). First order universality is also known when G is (possibly high dimensional) grid (e.g., $G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$) (Yarotsky, 2018), a case that includes periodic convolutional neural networks. Universality of first order networks is also known when $G = S_n$ (Zaheer et al., 2017; Qi et al., 2017; Yarotsky, 2018) in the context of invariant networks that operate on sets or point clouds.

Our goal in this section is to derive a necessary condition on G for the universality of first order G-invariant networks. To this end, we first find a function, playing the role of the Vandermonde polynomial in the previous section, that is G-invariant but not H-invariant, where $G < H \leq S_n$.

Lemma 3. *Let $G < H \leq S_n$. Then there exists a continuous function $f : \mathbb{R}^n \to \mathbb{R}$ which is G-invariant but not H-invariant.*

Proof. Pick a point $x_0 \in \mathbb{R}^n$ with distinct coordinates. Since the stabilizer $(S_n)_{x_0}$ is trivial (i.e., no permutation fixes x_0 excluding the identity), the size of the orbits of x_0 equals the size of the acting group. Namely, $|G \cdot x_0| = |G|$ and $|H \cdot x_0| = |H|$. Furthermore, since $|G| < |H|$ and $G \cdot x_0 \subset H \cdot x_0$, we get that the H orbit strictly includes the G orbit. That is, $G \cdot x \subsetneq H \cdot x$. Since $H \cdot x_0$ is a finite set of points, there exists a continuous function $\hat{f}$ such that $\hat{f}|_{G \cdot x_0} = 1$, and $\hat{f}|_{H \cdot x_0 \setminus G \cdot x_0} = 0$. Define $f(x) = \frac{1}{|G|} \sum_{g \in G} \hat{f}(g \cdot x)$. Now, f is G-invariant by construction but $f(x_0) = 1$ and $f(h \cdot x_0) = 0$ for $h \cdot x_0 \in H \cdot x_0 \setminus G \cdot x_0$. Therefore, f is not H-invariant. □

In case of first order G-invariant networks the equivariant/invariant layers have the form $L : \mathbb{R}^{n \times a} \to \mathbb{R}^{n \times b}$ and satisfy the fixed point equations (9). The solution set of the purely linear equivariant layers consists of tensors $\mathbf{L} \in \mathbb{R}^{n^2 \times a \times b}$ that are constant on equivalence classes of indices defined by the equivalence relation $(i_1, i_2) \sim (j_1, j_2)$ if there exists $g \in G$ so that $j_\ell = g(i_\ell)$, $\ell = 1, 2$. We denote the number of equivalence classes by $|[n]^2/G|$. The solution set of constant equivariant operators are tensors $\mathbf{B} \in \mathbb{R}^{n \times b}$ that are constant on equivalence classes defined

[4] S_n is in-fact n-transitive and is therefore also k-transitive for all $k \leq n$.

by the equivalence relation $i \sim j$ if there exists $g \in G$ so that $j = g(i)$. We denote the number of these classes by $|[n]/G|$. We prove:

Theorem 3. *Let $G \leq S_n$. If first order G-invariant networks are universal, then $\left|[n]^2/H\right| < \left|[n]^2/G\right|$ for any strict super-group $G < H \leq S_n$.*

Proof. Assume by contradiction that there exists a strict super-group $G < H \leq S_n$ so that $\left|[n]^2/G\right| = \left|[n]^2/H\right|$. This in particular means that $|[n]/G| = |[n]/H|$. Therefore $\mathcal{L}^G = \mathcal{L}^H$. That is, the spaces of equivariant and invariant linear layers coincide for G and H. This implies, as before, that every first order G-invariant network is also H-invariant.

We proceed similarly to the proof of theorem 2: By lemma 3, there exists a continuous function $f : \mathbb{R}^n \to \mathbb{R}$ that is G-invariant but not H-invariant. Let x_0 be a point with distinct coordinates where $f(x_0) = 1$ (it exists by construction, see proof of theorem 2). Furthermore, by construction $f(h \cdot x_0) = 0$ if $h \cdot x_0 \in H \cdot x_0 \setminus G \cdot x_0$.

Let $\epsilon > 0$ and $K \subset \mathbb{R}^n$ a compact set containing both $x_0, h \cdot x_0$. Assume by way of contradiction that there exists a first order G-invariant network F (which is also H-invariant in view of the above) such that $|f(x_0) - F(x_0)| \leq \epsilon$ as well as:

$$|f(h \cdot x_0) - F(h \cdot x_0)| = |f(h \cdot x_0) - F(x_0)| \leq \epsilon.$$

These last equations imply that $1 = |f(x_0) - f(h \cdot x_0)| \leq |f(x_0) - F(x_0)| + |F(x_0) - f(h \cdot x_0)| \leq 2\epsilon$ and since ϵ is arbitrary we get a contradiction. □

Using theorem 3 we can show that there exist a few infinite families of permutation groups (excluding the alternating group A_n) for which first order invariant networks are not universal. For example, any strict subgroup $G < S_n$ that is 2-transitive is such a group since in this case $\left|[n]^2/G\right| = \left|[n]^2/S_n\right|$ and consequently G-invariant/equivariant layers are also S_n-invariant/equivariant. Examples of 2-transitive permutation groups include projective linear groups over finite fields $PSL_d(F_q)$ (for $q = p^n$ where $p, n \in \mathbb{N}$, p is prime) that act on the finite projective space, and can be seen as a subgroup of S_n for $n = (q^d - 1)/(q - 1)$ (the number of elements in this space). Similarly affine subgroups over finite fields $A\Gamma L_d(F_q)$ that act on F_q^d can be shown to be 2-transitive as a subgroup of S_n for $n = q^d$. See (Dixon & Mortimer, 1996) for a full classification of 2-transitive subgroups of S_n.

Relation to (Ravanbakhsh et al., 2017). Groups for which the condition in theorem 3 holds are called 2-closed and were first introduced by (Wielandt, 1969) (see (Babai, 1995) for further study). Theorem 3 reveals an interesting connection between our work and the work of (Ravanbakhsh et al., 2017) that studies parameter sharing schemes. One of the basic notions defined in their paper is the notion of *uniquely G-equivariant functions*, which describes functions that are G-equivariant but not equivariant to any super-group of G. For example, a consequence of proposition 6 is that $A_n \leq S_n$ (with the representation used in this paper) has no uniquely equivariant linear functions between tensors of total order $\leq n - 2$. It was shown in (Ravanbakhsh et al., 2017) that 2-closed groups are exactly the groups for which one can find a uniquely equivariant function. In this section we proved that the existence of a uniquely G-equivariant linear function is a necessary condition for first order universality. As stated in (Ravanbakhsh et al., 2017) some examples for 2-closed groups are fixed-point free groups (e.g., the cyclic group C_n) and S_n itself.

6. Conclusion

In this paper we have considered the universal approximation property of a popular invariant neural network model. We have shown that these networks are universal with a construction that uses tensors of order $\leq \frac{n(n-1)}{2}$, which makes this architecture impractical. On the other hand, there exists a permutation group for which we have proved a lower bound of $\frac{n-2}{2}$ on the tensor order required to achieve universality. We then addressed the more practical question of which groups G allow first order G-invariant networks to be universal. We have proved that 2-closedness of G is a necessary condition, and gave examples of infinite permutation group families that do not satisfy this condition.

Our work is a first step in advancing the understanding of approximation power of a large class of invariant neural networks that becomes increasingly popular in applications. Several questions remain open: First, a classification of 2-closed groups will give us a complete answer to which networks are first-order universal. As far as we know this is an open question in group theory. Still, mapping the 2-closed landscape for specific groups G that are interesting for machine learning applications is a worthy challenge. Second, In case one wishes to construct a G-invariant network for a group G that is not 2-closed, developing fast and efficient implementations of higher order layers seems like a potentially useful direction. Lastly, another interesting venue for future work might be to come up with new, possibly non-linear, models for invariant networks.

Acknowledgments

This research was supported in part by the European Research Council (ERC Consolidator Grant, "LiftMatch" 771136) and the Israel Science Foundation (Grant No. 1830/17).

References

Babai, L. Automorphism groups, isomorphism, reconstruction. chapter 27 of the handbook of combinatorics, 1447–1540. rl graham, m. grötschel, l. lovász eds, 1995.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999, 2016a.

Cohen, T. S. and Welling, M. Steerable CNNs. (1990):1–14, 2016b. URL http://arxiv.org/abs/1612.08498.

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

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

Dixon, J. D. and Mortimer, B. *Permutation groups*, volume 163. Springer Science & Business Media, 1996.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272, 2017.

Göbel, M. Computing bases for rings of permutation-invariant polynomials. *J. Symb. Comput.*, 19(4):285–291, 1995.

Hartford, J., Graham, D. R., Leyton-Brown, K., and Ravanbakhsh, S. Deep models of interactions across sets. *arXiv preprint arXiv:1803.02879*, 2018.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

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.

Kraft, H. and Procesi, C. Classical invariant theory, a primer. *Lecture Notes, Version*, 2000.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Syx72jC9tm.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017.

Ravanbakhsh, S., Schneider, J., and Poczos, B. Equivariance through parameter-sharing. *arXiv preprint arXiv:1702.08389*, 2017.

Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. 2018. URL http://arxiv.org/abs/1807.02547.

Wielandt, H. Permutation groups through invariant relations and invariant functions. 1969.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

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*, pp. 3391–3401, 2017.