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. Lastly, we propose a conjecture stating that this condition is also sufficient.

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 $[n] = \{1, 2, \ldots, n\}$. 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.

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-cloud 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) 1.

1It is common to use convolutional layers without cyclic padding which implies that these networks are not precisely translation invariant.
We conclude the paper by considering the question: For which groups \( G \) does the action of \( G \) on tensors make use of linear equivariant layers between tensors \( \mathbb{R}^n \) and is applied to each dimension of the tensor. A transposition of two numbers is an invariant function.

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

\[
(g \cdot X)_{i_1 \ldots i_k, j} = X_{g^{-1}(i_1) \ldots g^{-1}(i_k), j}, \quad g \in G.
\]

The action of \( G \) on tensors \( X \in \mathbb{R}^{n \times a} \) is invariant if \( X \) is \( G \)-invariant, i.e., \( g \cdot X = X \) for all \( g \in G \). We prove: 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 for networks to have this property, give examples of infinite families of permutation groups that do not possess it, and propose a conjecture stating that our condition is also a sufficient condition.

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

\[
g \cdot x = (x_{g^{-1}(1)}, \ldots, x_{g^{-1}(n)}), \quad g \in G.
\]

The action of \( G \) on tensors \( X \in \mathbb{R}^{n \times a} \) is invariant if \( X \) is \( G \)-invariant, i.e., \( g \cdot X = X \) for all \( g \in G \). We prove: 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 for networks to have this property, give examples of infinite families of permutation groups that do not possess it, and propose a conjecture stating that our condition is also a sufficient condition.

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

\[
g \cdot x = (x_{g^{-1}(1)}, \ldots, x_{g^{-1}(n)}), \quad g \in G.
\]

The action of \( G \) on tensors \( X \in \mathbb{R}^{n \times a} \) is defined similarly by

\[
(g \cdot X)_{i_1 \ldots i_k, j} = X_{g^{-1}(i_1) \ldots g^{-1}(i_k), j}, \quad g \in G.
\]

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 \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) 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 \( L : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{n^2} \) satisfying \( L(g \cdot x) = g \cdot L(x) \), for all \( g \in G \), and \( x \in \mathbb{R}^{n^2} \). An invariant linear layer is an affine map \( h : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^b \) satisfying \( h(g \cdot x) = h(x) \), for all \( g \in G \), and \( x \in \mathbb{R}^{n^2} \).

A common way to construct \( G \)-invariant networks is:

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

\[
F = m \circ h \circ L_d \circ \sigma \circ \cdots \circ \sigma \circ L_1,
\]

where \( L_i \) are linear \( G \)-equivariant layers, \( \sigma \) is an activation function,

\[
\text{or equivalently } \sigma \circ \cdots \circ \sigma \circ L_1,
\]

**Theorem 1.** Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) 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.

(\( \text{Yarotsky}, 2018 \)). However, universality is not known for tabular and multi-set networks (\( \text{Hartford et al.}, 2018 \)), graph and hyper-graph networks (\( \text{Kondor et al.}, 2018; \text{Maron et al.}, 2019 \)); 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 (\( \text{Yarotsky}, 2018 \)) 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 \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is approximated using a \( G \)-invariant network. The proof is constructive and suggests use of linear equivariant layers between tensors \( \mathbb{R}^n \) and is applied to each dimension of the tensor. Therefore, an invariant function.

\[
\mathbb{R}^n \rightarrow \mathbb{R} \text{ that satisfies } f(g \cdot x) = f(x) \text{ for all } x \in \mathbb{R}^n \text{ and } g \in G.
\]

**Definition 2.** A linear equivariant layer is an affine map \( L : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^{n^2} \) satisfying \( L(g \cdot x) = g \cdot L(x) \), for all \( g \in G \), and \( x \in \mathbb{R}^{n^2} \). An invariant linear layer is an affine map \( h : \mathbb{R}^{n^2} \rightarrow \mathbb{R}^b \) satisfying \( h(g \cdot x) = h(x) \), for all \( g \in G \), and \( x \in \mathbb{R}^{n^2} \).

A common way to construct \( G \)-invariant networks is:

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

\[
F = m \circ h \circ L_d \circ \sigma \circ \cdots \circ \sigma \circ L_1,
\]

where \( L_i \) are linear \( G \)-equivariant layers, \( \sigma \) is an activation function,

\[
\text{or equivalently } \sigma \circ \cdots \circ \sigma \circ L_1,
\]

**Theorem 1.** Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) 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.

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 (MLP).
\( X \in \mathbb{R}^d \) 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 \( G \)-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 \( \frac{|[n]|^2/H}{|n|^2} < \frac{|[n]|^2/G}{|n|^2} \) for any strict super-group \( G < H \leq S_n \).

\( \frac{|[n]|^2/G}{|n|^2} \) is the number of equivalence classes of \( [n] \) 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 supergroups of \( G \) have strictly better separation of the double index space \( [n] \). We conclude the paper with the conjecture that this condition is also sufficient.

### 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 \( x_1, \ldots, x_n \) over \( \mathbb{R} \) that are also \( G \)-invariant functions. They are denoted \( \mathbb{R}[x_1, \ldots, x_n]^G \), where \( \mathbb{R}[x_1, \ldots, x_n] \) is the set of all polynomials over \( \mathbb{R} \).

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

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

\( q \) is a \( G \)-invariant polynomial and hence

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

furthermore for \( x \in K' \):

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

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 \( p \in \mathbb{R}[x_1, \ldots, x_n]^G \), there exists a \( G \)-invariant network \( F \) that approximates \( p \) to an \( \epsilon \)-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., \( p_k \in \mathbb{R}[x_1, \ldots, x_n]^G \).

(ii) We characterize all homogeneous \( G \)-invariant polynomials of a fixed degree \( k \). In particular we find a basis to all such polynomials, \( b_{k1}, b_{k2}, \ldots, b_{kn_k} \in \mathbb{R}[x_1, \ldots, x_n]^G \). 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 \( \epsilon \)-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 \( p : \mathbb{R}^n \rightarrow \mathbb{R} \) 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 $k = 0, 1, \ldots, d$, i.e., $\mathbb{R}_k[x_1, \ldots, x_n]^G$. Any homogeneous polynomial of degree $k$ can be written as

$$p(x) = \sum_{i_1, \ldots, i_k=1}^n W_{i_1 \ldots i_k} x_{i_1} \cdots x_{i_k}, \quad (4)$$

where $W \in \mathbb{R}^{n^k}$ 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 $W$ can be obtained by taking a symmetric matrix. That is, $W$ that satisfies $W_{i_1 \ldots i_k} = W_{\sigma(i_1) \cdots \sigma(i_k)}$, for all $\sigma \in S_k$. In short, we ask $W \in \text{Sym}^k(\mathbb{R}^n) \subset \mathbb{R}^{n^k}$. For example, the case $k = 2$ amounts to representing a quadratic form using a symmetric matrix. 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 $W \in \mathbb{R}^{n^k}$:

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

$$g \cdot W = W, \quad \forall g \in G. \quad (5)$$

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

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

$$= \sum_{i_1, \ldots, i_k=1}^n W_{g(i_1) \ldots g(i_k)} x_{i_1} \cdots x_{i_k}.$$  

By equating monomials’ coefficients of $p(x)$ and $p(g \cdot x)$ and the symmetry of $W$ we get

$$W_{i_1 \ldots i_k} = W_{g(i_1) \ldots g(i_k)}.$$  

This implies that $W$ satisfies $g \cdot W = W$ for all $g \in G$.  

Equation 5 is a linear homogeneous system of equations and therefore the set of solutions $W$ 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 $j_k = g(\sigma(i_k))$, $\ell = 1, \ldots, k$. Intuitively, $g$ takes care of the $G$-invariance while $\sigma$ 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 $\tau$ and called the $k$-classes. We show:

**Proposition 3.** The set of polynomials

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

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

Proof. Denote $W^\tau$ the symmetric coefficient tensor of $p^\tau$. Note that

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

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 $p^\tau$, with $\tau$ a $k$-classes, is a linearly independent set since each $p^\tau$ contains a different collection of monomials. By Proposition 2, the symmetric coefficient tensor $W$ of every $g \in \mathbb{R}_k[x_1, \ldots, x_n]^G$ satisfies the fixed-point equation, equation 5. This in particular means that $W$ is constant on its $k$-classes. Hence $W$ can be written as linear combination of $W^\tau$, 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 $W$ which are needed since every polynomial of degree $> 1$ has several representing tensors $W$.

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

**Proposition 4.** Let $\tau$ 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 $x \in \mathbb{R}^n$

$$L^\tau_\ell(x)_{i_1 \ldots i_k} = \begin{cases} x_{i_k} & (i_1, \ldots, i_k) \in \tau \\ 0 & \text{otherwise} \end{cases}. \quad (8)$$

Then $L^\tau_\ell$ is a linear $G$-equivariant function, that is

$$L^\tau_\ell(g \cdot x) = g \cdot L^\tau_\ell(x), \quad \forall x \in \mathbb{R}^n, g \in G. \quad (9)$$

**Proof.** We have:

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

On the other hand,

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

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 $\tau$.  

\[
W_{i_1 \ldots i_k} = \begin{cases} 1 & (i_1, \ldots, i_k) \in \tau \\ 0 & \text{otherwise} \end{cases}. \quad (7)
\]
Next, we construct the approximating $G$-invariant network:

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

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

Consider the following $G$-invariant network: First, given an input $x \in \mathbb{R}^n$ map it to $\mathbb{R}^{n^{k+1}}$ (i.e., $k$ is the number of channels) by

$$L^\tau(x)_{i_1 \ldots i_k, \ell} = L^\tau_i(x)_{i_1 \ldots i_k},$$

where in the last inequality we used the $n$-2017)).

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

**Proof.** (proposition 1) Let us denote by $b_{k_1, \ldots, k_n}$ the polynomials $p^\tau$, with $\tau$ the $k$-classes. Let $F^{k_j}$ denote the $G$-invariant network approximating $b_{k_j}$, $k = 0,1,\ldots, d$, to an $\epsilon$-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

$$\left| p(x) - \sum_{k=0}^d \sum_{j=1}^{n_k} \alpha_{kj} F^{k_j}(x) \right| \leq \epsilon \|\alpha\|_1,$$

where $\|\alpha\|_1 = \sum_{k,j} |\alpha_{kj}|$ depends only upon $p$, where $\epsilon$ is arbitrary. To finish the proof we need to show that $F = \sum_{k=0}^d \sum_{j=1}^{n_k} \alpha_{kj} F^{k_j}$ 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 appendix:

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

### 3.2. Bounded order construction

We have constructed a $G$-invariant network $F$ that approximates an arbitrary $G$-invariant polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]^G$ 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 $G \leq S_n$. 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 $f_1, \ldots, f_m \in \mathbb{R}[x_1, \ldots, x_n]^G$ such that any invariant polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]^G$ can be expressed as

$$p(x) = h(f_1(x), \ldots, f_m(x)),$$

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

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 $G \leq S_n$, there exists a generating set of $G$-invariant polynomials of degree bounded by $\frac{n(n-1)}{2}$, for $n \geq 3$, see (Göbel, 1995). We can now repeat the construction above, building a $G$-invariant network $F$, approximating $f_i$ to a $\epsilon_i$-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 $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with the final output approximating $f(x) = (f_1(x), \ldots, f_m(x))$ to a $\epsilon$-accuracy. Now we compose the output of $F$ with an MLP $H : \mathbb{R}^m \rightarrow \mathbb{R}$ approximating the polynomial $h$ over the compact set $f(K) + B_\epsilon \subset \mathbb{R}^m$ to an $\epsilon$-accuracy, $B_\epsilon$ is a closed ball centered at the origin of radius $\epsilon$ 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:

$$|H(F(x)) - h(f(x))| \leq |H(F(x)) - H(f(x))| + |H(f(x)) - h(f(x))| \leq 2\epsilon,$$

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. Example: Universality of (hyper-) graph networks

Graph, or hyper-graph data can be described using tensors $X \in \mathbb{R}^{n^k \times a}$, 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 $X \in \mathbb{R}^{n^2}$. The graph symmetries are reordering the vertices by a permutation, namely $g \cdot X$, 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 proposition 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).

### 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 $L : \mathbb{R}^{n^k \times a} \rightarrow \mathbb{R}^{n^l \times b}$, where $l = 0$ represents the invariant case. By definition $L(g \cdot X) = g \cdot L(X)$ for all $X \in \mathbb{R}^{n^k \times a}$. In particular this means that

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

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 $L \in \mathbb{R}^{n^{k+l} \times a \times b}$ these equations become the fixed-point equation for linear equivariant layers (see appendix for derivation): $g \cdot L = L$, $g \in G$. (9)

The constant part of $L$ can be encoded using a tensor $B \in \mathbb{R}^{n^l \times b}$ 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 $L^G$ the collection of $L : \mathbb{R}^{n^k \times a} \rightarrow \mathbb{R}^{n^l \times b}$ linear $G$-equivariant ($l > 0$) or $G$-invariant ($l = 0$) layers.

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

**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 $G = S_n$. (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.)
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. 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 \( i_\alpha = j_\beta \) 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 argument applies to the constant part, as well. \( \square \)

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:

\[
|V(g \cdot x) - F(g \cdot x)| = |(-1)^{\ell} V(x) - F(x)| = |V(x) + F(x)| \leq \epsilon
\]

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

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 and propose a conjecture. 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 universality approximation theorems (Cybenko, 1989; Hornik, 1991). First order universality is also known when \( G \) is possibly high dimensional (grid, e.g., \( G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_m} \)) (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, \( G \cdot x \nsubseteq H \cdot x \). Since \( H \cdot x_0 \) is a finite set of points, there exists a continuous function \( f \) such that \( f|_{G \cdot x_0} = 1 \), and \( f|_{H \cdot x_0 \setminus G \cdot x_0} = 0 \). Define \( f(x) = \frac{1}{|G|} \sum_{g \in G} f(g \cdot x) \). 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. \( \square \)

In case of first order \( G \)-invariant networks the equivariant/linear 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 \( \mathbf{L} \in \mathbb{R}^{n \times a \times b} \) 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 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 \(|\{n\}^2/G| < |\{n\}^2/G| \) 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 $\left| [n] / G \right| = \left| [n] / H \right|$. Therefore $L^G = L^H$. 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 $\epsilon$ is arbitrary we get a contradiction. \qed

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 $AGL_d(F_q)$ 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.

We conclude this section with a conjecture that 2-closedness is not only a necessary condition for universality of first-order invariant networks, but also sufficient:

Conjecture 1. Let $G \leq S_n$. First order $G$-invariant networks are universal if and only if $G$ is 2-closed.

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, gave examples of infinite permutation group families that do not satisfy this condition, and conjectured this condition is also sufficient.

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.

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
On the Universality of Invariant Networks


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


A. Proofs

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

Proof. We need to show that $F = \sum_{k=0}^{d} \sum_{j=1}^{\tau} F^{kj}$ can indeed be realized as a single, unified $G$-invariant network. As we already saw, each network $F^{kj}$ has the structure

$$\mathbb{R}^{n} \xrightarrow{L} \mathbb{R}^{n \times k} \xrightarrow{M} \mathbb{R}^{n} \rightarrow \mathbb{R},$$

with a suitable $k$-class $\tau$. To create the unified $G$-invariant network we first lift each $F^{kj}$ to the maximal dimension $d$. That is, $\tilde{F}^{kj}$ with the structure

$$\mathbb{R}^{n} \xrightarrow{\tilde{L}^{kj}} \mathbb{R}^{n \times k} \xrightarrow{\tilde{M}^{k}} \mathbb{R}^{n} \rightarrow \mathbb{R}.$$ 

This is done by composing each equivariant layer $L : \mathbb{R}^{n \times a} \rightarrow \mathbb{R}^{n \times b}$ with two linear equivariant operators $U^{b} : \mathbb{R}^{n \times b} \rightarrow \mathbb{R}^{n \times b}$ and $D^{a} : \mathbb{R}^{d \times a} \rightarrow \mathbb{R}^{n \times a}$,

$$U^{b}L D^{a} : \mathbb{R}^{d \times a} \rightarrow \mathbb{R}^{n \times b},$$

where

$$U^{b}(x)_{1 \ldots i_{d}, j} = x_{i_{1} \ldots i_{d}j}$$

and

$$D^{a}(y)_{i_{1} \ldots i_{k}, j} = n^{k-d} \sum_{i_{k+1} \ldots i_{d}=1}^{n} y_{i_{k+1} \ldots i_{d}j}.$$ 

Since $U^{b}, D^{a}$ are equivariant, $U^{b}L D^{a}$ in equation 10 is equivariant. Furthermore $D^{a} \circ \sigma \circ U^{a} = \sigma$, where $\sigma$ is the pointwise activation function. Lastly, given two $G$-invariant networks with the same tensor order $d$ they can be combined to a single $G$-invariant network by concatenating their features. That is, if $L_{1} : \mathbb{R}^{d \times a} \rightarrow \mathbb{R}^{d \times b}$, and $L_{2} : \mathbb{R}^{d \times a'} \rightarrow \mathbb{R}^{d \times b'}$, then their concatenation would yield $L_{1,2} : \mathbb{R}^{d \times (a+a')} \rightarrow \mathbb{R}^{d \times (b+b')}$. Applying this concatenation to all $\tilde{F}^{kj}$ we get our unified $G$-invariant network.

Writing equation 11 using this notation gives:

$$\sum_{i_{1} \ldots i_{k}, i} L_{g(j_{1}) \ldots g(j_{i}), i_{1} \ldots i_{k}, i, j} X_{g^{-1}(i_{1}) \ldots g^{-1}(i_{k}), i}$$

$$= \sum_{i_{1} \ldots i_{k}, i} L_{g(j_{1}) \ldots g(j_{i}), g(i_{1}) \ldots g(i_{k}), i, j} X_{i_{1} \ldots i_{k}, i}$$

for all $g \in G$ and $X \in \mathbb{R}^{n \times a}$. This implies equation 9, namely

$$g \cdot L = L, \ g \in G.$$

The constant part of $L$ is done similarly.