INARIANT AND EQUIVARIANT GRAPH NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Invariant and equivariant networks have been successfully used for learning images, sets, point clouds, and graphs. A basic challenge in developing such networks is finding the maximal collection of invariant and equivariant linear layers. Although this question is answered for the first three examples (for popular transformations, at-least), a full characterization of invariant and equivariant linear layers for graphs is not known.

In this paper we provide such a characterization and show that the dimension of invariant and equivariant linear layers between edge-value graph data is 2 and 15, respectively. More generally, for graph data defined on \( k \)-tuples of nodes, the dimension is the \( k \)-th and \( 2k \)-th Bell numbers. Orthogonal bases for the layers are computed, including generalization to multi-graph data. The constant number of basis elements and their characteristics allow successfully applying the networks to different size graphs. From the theoretical point of view, our results generalize and unify recent advancement in equivariant deep learning.

Applying these new linear layers in a simple deep neural network framework is shown to achieve comparable results to state-of-the-art and to have better expressivity than previous invariant and equivariant bases.

1 INTRODUCTION

We consider the problem of graph learning, namely finding a functional relation between input graphs (more generally, hyper-graphs) \( \mathcal{G}^\ell \) and corresponding targets \( T^\ell \), e.g., labels. As graphs are common data representations, this task received quite a bit of recent attention in the machine learning community [Bruna et al. (2013); Henaff et al. (2015); Monti et al. (2017); Ying et al. (2018)].

More specifically, a (hyper-)graph data point \( \mathcal{G} = (\mathbb{V}, \mathbb{A}) \) consists of a set of \( n \) nodes \( \mathbb{V} \), and values \( \mathbb{A} \) attached to its hyper-edge\(^1\). These values are encoded as a tensor \( \mathbb{A} \). The order of the tensor \( \mathbb{A} \), or equivalently, the number of indices used to represent its elements, indicates the type of data it represents, as follows: First order tensor represents node-values where \( \mathbb{A}_i \) is the value of the \( i \)-th node; Second order tensor represents edge-values, where \( \mathbb{A}_{ij} \) is the value attached to the \( (i, j) \) edge; in general, \( k \)-th order tensor encodes hyper-edge-values, where \( \mathbb{A}_{i_1, \ldots, i_k} \) represents the value of the hyper-edge represented by \( (i_1, \ldots, i_k) \). For example, it is customary to represent a graph using a binary adjacency matrix \( \mathbb{A} \), where \( \mathbb{A}_{ij} \) equals one if vertex \( i \) is connected to vertex \( j \) and zero otherwise. We denote the set of order-\( k \) tensors by \( \mathbb{R}^{n^k} \).

The task at hand is constructing a functional relation \( f(\mathbb{A}^\ell) \approx T^\ell \), where \( f \) is a neural network. If \( T^\ell = t^\ell \) is a single output response then it is natural to ask that \( f \) is order invariant, namely it should produce the same output regardless of the node numbering used to encode \( \mathbb{A} \). For example, if we represent a graph using an adjacency matrix \( \mathbb{A} = \mathbb{A} \in \mathbb{R}^{n \times n} \), then for an arbitrary permutation matrix \( \mathbb{P} \) and an arbitrary adjacency matrix \( \mathbb{A} \), the function \( f \) is order invariant if it satisfies \( f(\mathbb{P}^T \mathbb{A} \mathbb{P}) = f(\mathbb{A}) \). If the targets \( T^\ell \) specify output response in a form of a tensor, \( T^\ell = T^\ell \), then it is natural to ask that \( f \) is order equivariant, that is, \( f \) commutes with the renumbering of nodes operator acting on tensors. Using the above adjacency matrix example, for every adjacency matrix \( \mathbb{A} \) and every permutation matrix \( \mathbb{P} \), the function \( f \) is equivariant if it satisfies \( f(\mathbb{P}^T \mathbb{A} \mathbb{P}) = \mathbb{P}^T f(\mathbb{A}) \mathbb{P} \).

To define invariance and equivariance for functions acting on general tensors \( \mathbb{A} \in \mathbb{R}^{n^k} \) we use the reordering operator: \( \mathbb{P} \approx \mathbb{A} \) is defined to be the tensor that results from renumbering the nodes \( \mathbb{V} \)

\(^1\)A hyper-edge is an ordered subset of the nodes, \( \mathbb{V} \)
Our analysis allows generalizing the multi-node set case to arbitrary tensor data over $\mathbb{V}_1 \times \mathbb{V}_2 \times \cdots \times \mathbb{V}_m$. Namely, for data points in the form of a tensor $A \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m}$, the tensor $A$ attaches a value to every element of the cartesian product $\mathbb{V}_1 \times \cdots \times \mathbb{V}_m$. The tensor $A$ attaches a value to every element of the cartesian product $\mathbb{V}_1, k_2$-tuple from $\mathbb{V}_2$, and so forth. We show that the linear space of invariant linear layers $L : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m} \to \mathbb{R}$ is of dimension $\prod_{i=1}^m b(k_i)$, while the equivariant linear layers $L : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m} \to \mathbb{R}^{n'_1 \times n'_2 \times \cdots \times n'_{m'}}$ has dimension $\prod_{i=1}^{m'} b(k_i + l_i)$. We also provide orthogonal bases for these spaces. Note that, for clarity, the discussion above disregards biases and features; we detail these in the paper.
In the experimental part of the paper we concentrated on possibly the most popular instantiation of graph learning, namely that of a single node set and edge-value data, e.g., with adjacency matrices. We created simple networks by composing our invariant or equivariant linear layers in standard ways and tested the networks in learning invariant and equivariant graph functions: (i) We compared identical networks with our bases and the bases of [Hartford et al., 2018] and showed we can learn graph functions like trace, diagonal, and maximal singular vector, while [Hartford et al., 2018] fail on these tasks, demonstrating their bases are not maximal for the graph-learning scenario. We also demonstrate our representation allows extrapolation: learning on one size graphs and testing on another size; (ii) We also tested our networks on a collection of graph learning datasets, achieving state-of-the-art results in 3 social network datasets.

2 PREVIOUS WORK

Our work builds on two main sub-fields of deep learning: group invariant or equivariant networks, and deep learning on graphs. Here we briefly review the relevant works.

**Invariance and equivariance in deep learning.** In many learning tasks the functions that we want to learn are invariant or equivariant to certain symmetries of the input object description. Maybe the first example is the celebrated translation invariance of Convolutional Neural Networks (CNNs) (LeCun et al., 1989; Krizhevsky et al., 2012); in this case, the image label is invariant to a translation of the input image. In recent years this idea was generalized to other types of symmetries such as rotational symmetries (Cohen & Welling, 2016a;b, Weiler et al., 2018; Cohen et al., 2018). Cohen & Welling (2016a) introduced Group Equivariant Neural Networks that use a generalization of the convolution operator to groups of rotations and reflections; Weiler et al., 2018; Cohen et al., 2018 also considered rotational symmetries but in the case of 3D shapes and spherical functions. Kavan-bakhsh et al. (2017) showed that any equivariant layer is equivalent to a certain parameter sharing scheme. If we adopt this point of view, our work reveals the structure of the parameter sharing in the case of graphs and hyper-graphs. In another work, Kondor & Trivedi (2018) show that a neural network layer is equivariant to the action of some compact group iff it implements a generalized form of the convolution operator. Yarotsky (2018) suggested certain group invariant/equivariant models and proved their universality. To the best of our knowledge these models were not implemented.

**Learning of graphs.** Learning of graphs is of huge interest in machine learning and we restrict our attention to recent advancements in deep learning on graphs. Gori et al. (2005); Scarselli et al. (2009) introduced Graph Neural Networks (GNN): GNNs hold a state (a real valued vector) for each node in the graph, and propagate these states according to the graph structure and learned parametric functions. This idea was further developed in [Li et al., 2015] that use gated recurrent units. Following the success of CNNs, numerous works suggested ways to define convolution operator on graphs. One promising approach is to define convolution by imitating its spectral properties using the Laplacian operator to define generalized Fourier basis on graphs (Bruna et al., 2013). Multiple follow-up works [Henaff et al., 2015; Defferrard et al., 2016; Kipf & Welling, 2016; Levie et al., 2017] suggest more efficient and spatially localized filters. The main drawback of spectral approaches is that the generalized Fourier basis is graph-dependent and applying the same network to different graphs can be challenging. Another popular way to generalize the convolution operator to graphs is learning stationary functions that operate on neighbors of each node and update its current state ([Atwood & Towsley, 2016; Duvendau et al., 2015; Hamilton et al., 2017; Niepert et al., 2016; Velickovic et al., 2017; Monti et al., 2017, Simonovsky & Komodakis, 2017]. This idea generalizes the locality and weight sharing properties of the standard convolution operators on regular grids. As shown in the wonderful work of Gilmer et al. (2017), most of the the above mentioned methods (including the spectral methods) can be seen as instances of the general class of Message Passing Neural Networks.

3 LINEAR IN Variant AND E Quivariant LAYERS

In this section we characterize the collection of linear invariant and equivariant layers. We start with the case of a single node set $V$ of size $n$ and edge-value data, that is order 2 tensors $A = A \in \mathbb{R}^{n \times n}$. As a typical example imagine, as above, an adjacency matrix of a graph. We set a bit of notation. Given a matrix $X \in \mathbb{R}^{n \times b}$ we denote $\text{vec}(X) \in \mathbb{R}^{nb \times 1}$ its column stack, and by brackets the inverse action of reshaping to a square matrix, namely $[\text{vec}(X)] = X$. Let $P$ denote an arbitrary permutation and $P$ its corresponding permutation matrix.
Let $L \in \mathbb{R}^{t \times n^2}$ denote the matrix representing a general linear operator $L : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ in the standard basis, then $L$ is order invariant iff $L\text{vec}(P^T A P) = L\text{vec}(A)$. Using the property of the Kronecker product that $\text{vec}(XAY) = Y^T \otimes X\text{vec}(A)$, we get the equivalent equality $LP^T \otimes P^T \text{vec}(A) = L\text{vec}(A)$. Since the latter equality should hold for every $A$ we get (after transposing both sides of the equation) that order invariant $L$ is equivalent to the equation

$$P \otimes P \text{vec}(L) = \text{vec}(L) \quad (1)$$

for every permutation matrix $P$. Note that we used $L^T = \text{vec}(L)$.

For equivariant layers we consider a general linear operator $P : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ and its corresponding matrix $L \in \mathbb{R}^{n^2 \times n^2}$. Equivariance of $L$ is now equivalent to $[L\text{vec}(P^T A P)] = P^T [L\text{vec}(A)] P$. Using the above property of the Kronecker product again we get $LP^T \otimes P^T \text{vec}(A) = P^T \otimes P^T L\text{vec}(A)$. Noting that $P^T \otimes P^T$ is an $n^2 \times n^2$ permutation matrix and its inverse is $P \otimes P$ we get the equivalent equality $P \otimes P L P^T \otimes P^T \text{vec}(A) = L\text{vec}(A)$. As before, since this holds for every $A$ and using the properties of the Kronecker product we get that $L$ is order equivariant iff for all permutation matrices $P$

$$P \otimes P \otimes P \otimes P \text{vec}(L) = \text{vec}(L). \quad (2)$$

From equations [1] and [2] we see that finding invariant and equivariant linear layers for the order-2 tensor data over one node set requires finding fixed points of the permutation matrix group represented by Kronecker powers $P \otimes P \otimes \cdots \otimes P$ of permutation matrices $P$. As we show next, this is also the general case for order-$k$ tensor data $A \in \mathbb{R}^{n^k}$ over one node set, $\mathbb{V}$. That is,

$$\text{invariant } L : \quad P \otimes^k \text{vec}(L) = \text{vec}(L) \quad (3)$$

$$\text{equivariant } L : \quad P \otimes^{2k} \text{vec}(L) = \text{vec}(L) \quad (4)$$

for every permutation matrix $P$, where $P \otimes^k = \bigotimes_{i=1}^k P$. In equation 3, $L \in \mathbb{R}^{1 \times n^k}$ is the matrix of an invariant operator; and in equation 4, $L \in \mathbb{R}^{n^k \times n^k}$ is the matrix of an equivariant operator. We call equations [3,4] the fixed-point equations.

To see this, let us add a bit of notation first. Let $p$ denote the permutation corresponding to the permutation matrix $P$. We let $P \ast A$ denote the tensor that results from expressing the tensor $A$ after renumbering the nodes in $\mathbb{V}$ according to permutation $P$. Explicitly, the $(p(i_1), p(i_2), \ldots, p(i_k))$-th entry of $P \ast A$ equals the $(i_1, i_2, \ldots, i_k)$-th entry of $A$. The matrix that corresponds to the operator $P \ast$ in the standard tensor basis $e_{i_1} \otimes \cdots \otimes e_{i_k}$ is the Kronecker power $P \otimes^k = (P^T)^{\otimes k}$. Note that vec($A$) is exactly the coordinate vector of the tensor $A$ in this standard basis and therefore we have vec($P \ast A$) = $P \otimes^k \text{vec}(A)$. We now show:

**Proposition 1.** A linear layer is invariant (equivariant) if and only if its coefficient matrix satisfies the fixed-point equations, namely equation [3] (equation [4]).

**Proof.** Similarly to the argument from the order-2 case, let $L \in \mathbb{R}^{1 \times n^k}$ denote the matrix corresponding to a general linear operator $L : \mathbb{R}^{n^k} \rightarrow \mathbb{R}$. Order invariance means

$$L\text{vec}(P \ast A) = L\text{vec}(A). \quad (5)$$

Using the matrix $P \otimes^k$ we have equivalently $LP \otimes^k \text{vec}(A) = L\text{vec}(A)$ which is in turn equivalent to $P \otimes^k \text{vec}(L) = \text{vec}(L)$ for all permutation matrices $P$. For order equivariance, let $L \in \mathbb{R}^{n^k \times n^k}$ denote the matrix of a general linear operator $L : \mathbb{R}^{n^k} \rightarrow \mathbb{R}^{n^k}$. Now equivariance of $L$ is equivalent to

$$[L\text{vec}(P \ast A)] = P \ast [L\text{vec}(A)]. \quad (6)$$

Similarly to above this is equivalent to $LP \otimes^k \text{vec}(A) = P \otimes^k L\text{vec}(A)$ which in turn leads to $P \otimes^k LP \otimes^k = L$, and using the Kronecker product properties we get $P \otimes^{2k} \text{vec}(L) = \text{vec}(L)$. \n
\n
### 3.1 Solving the fixed-point equations

We have reduced the problem of finding all invariant and equivariant linear operators $L$ to finding all solutions $L$ of equations [3] and [4]. Although the fixed point equations consist of an exponential number of equations with only a polynomial number of unknowns they actually possess a solution space of constant dimension (i.e., independent of $n$).
To find the solution of $P^{\otimes \ell} \text{vec}(X) = \text{vec}(X)$, where $X \in \mathbb{R}^{n^{\ell}}$, note that $P^{\otimes \ell} \text{vec}(X) = \text{vec}(Q \ast X)$, where $Q = P^{T}$. As above, the tensor $Q \ast X$ is the tensor resulted from renumbering the nodes in $\mathbb{V}$ using permutation $Q$. Equivalently, the fixed-point equations we need to solve can be formulated as 

$$Q \ast X = X, \quad \forall Q \text{ permutation matrices} \quad (7)$$

The permutation group is acting on tensors $X \in \mathbb{R}^{n^{\ell}}$ with the action $X \mapsto Q \ast X$. We are looking for fixed points under this action. To that end, let us define an equivalence relation in the index space of tensors $\mathbb{R}^{n^{\ell}}$, namely in $[n]^{\ell}$, where with a slight abuse of notation (we use light brackets) we set $[n] = \{1, 2, \ldots, n\}$. For multi-indices $a, b \in [n]^{\ell}$ we set $a \sim b$ iff $a, b$ have the same equality pattern, that is $a_i = a_j \Leftrightarrow b_i = b_j$ for all $i, j \in [\ell]$.

The equality pattern equivalence relation partitions the index set $[n]^{\ell}$ into equivalence classes, the collection of which is denoted $[n]^{\ell}/\sim$. Each equivalence class can be represented by a unique partition of the set $[\ell]$ where each set in the partition indicates maximal set of identical values. Let us exemplify. For $\ell = 2$ we have two equivalence classes $\gamma_1 = \{\{1\}, \{2\}\}$ and $\gamma_2 = \{\{1, 2\}\}$; $\gamma_1$ represents all multi-indices $(i, j)$ where $i \neq j$, while $\gamma_2$ represents all multi-indices $(i, j)$ where $i = j$. For $\ell = 4$, there are 15 equivalence classes $\gamma_1 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_2 = \{\{1\}, \{2\}, \{3, 4\}\}$, $\gamma_3 = \{\{1, 2\}, \{3\}, \{4\}\}$, $\gamma_4 = \{\{1\}, \{2\}, \{3, 4\}\}$, $\gamma_5 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_6 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_7 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_8 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_9 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_{10} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_{11} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_{12} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_{13} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_{14} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$, $\gamma_{15} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$.

For each equivalence class $\gamma \in [n]^{\ell}/\sim$ we define an order-$\ell$ tensor $B^\gamma \in \mathbb{R}^{n^{\ell}}$ by setting

$$B^\gamma_a = \begin{cases} 1 & a \in \gamma \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Since we have a tensor $B^\gamma$ for every equivalence class $\gamma$, and the equivalence classes are in one-to-one correspondence with partitions of the set $[\ell]$ we have $b(\ell)$ tensors $B^\gamma$. (Remember that $b(\ell)$ denotes the $\ell$-th Bell number.) We next prove:

**Proposition 2.** The tensors $B^\gamma$ in equation (8) form an orthogonal basis (in the standard inner-product) to the solution set of equations (7). The dimension of the solution set is therefore $b(\ell)$.

**Proof.** Let us first show that: $X$ is a solution to equation (7) iff $X$ is constant on equivalence classes of the equality pattern relation, $\sim$. Since permutation $q : [n] \to [n]$ is a bijection the equality pattern of $a = (i_1, i_2, \ldots, i_\ell) \in [n]^{\ell}$ and $q(a) = (q(i_1), q(i_2), \ldots, q(i_\ell)) \in [n]^{\ell}$ are identical, i.e., $a \sim q(a)$. Taking the $a \in [n]^{\ell}$ entry of both sides of equation (7) gives $X_{q(a)} = X_a$. Now, if $X$ is constant on equivalence classes then in particular it will have the same value at $a$ and $q(a)$ for all $a \in [n]^{\ell}$ and permutations $q$. Therefore $X$ is a solution to equation (7). For the only if part, consider a tensor $X$ for which there exist multi-indices $a \sim b$ (with identical equality patterns) and $X_a \neq X_b$ then $X$ is not a solution to equation (7). Indeed, since $a \sim b$ one can find a permutation $q$ so that $b = q(a)$ and using the equation above, $X_b = X_{q(a)} = X_a$ which leads to a contradiction.

To finish the proof note that any tensor $X$, constant on equivalence classes, can be written as a linear combination of $B^\gamma$, which are merely indicators of the equivalence class. Furthermore, the collection $B^\gamma$ have pairwise disjoint supports and therefore are an orthogonal basis. $\square$

Combining propositions (1) and (2) we get the characterization of invariant and equivariant linear layers acting on general $k$-order tensor data over a single node set $\mathbb{V}$:

**Theorem 1.** The space of invariant (equivariant) linear layers $\mathbb{R}^{n^{k}} \to \mathbb{R}$ ($\mathbb{R}^{n^{k}} \to \mathbb{R}^{n^{k}}$) is of dimension $b(k)$ ($b(2k)$) with basis elements $B^\gamma$ defined in equation (8) where $\gamma$ are equivalence classes in $[n]^{\ell}/\sim$, $([n]^{2k})/\sim$.

**Biases** Theorem 1 deals with purely linear layers, that is without bias, i.e., without constant part. Nevertheless extending the previous analysis to constant layers is straightforward. First, any constant layer $\mathbb{R}^{n^{k}} \to \mathbb{R}$ is also invariant so all constant invariant layers are represented by constants $c \in \mathbb{R}$. For equivariant layers $L : \mathbb{R}^{n^{k}} \to \mathbb{R}^{n^{k}}$ we note that equivariance means $C = L(P \ast A) = P \ast L(A) = P \ast C$. Representing this equation in matrix form we get $P^{T} \otimes \text{vec}(C) = \text{vec}(C)$. This shows that constant equivariant layers on one node set acting on general $k$-order tensors are also characterized by the fixed-point equations, and in fact have the same form and dimensionality as invariant layers on $k$-order tensors, see equation (3). Specifically, their basis is $B^\lambda$, $\lambda \in [n]^{k}/\sim$. For example, for $k = 2$, the biases are shown on the right in figure 4.
Features. It is pretty common that input tensors have vector values (i.e., features) attached to each hyper-edge (\(k\)-tuple of nodes) in \(\mathcal{V}\), that is \(\mathbf{A} \in \mathbb{R}^{n_k \times d}\). Now linear invariant \(\mathbb{R}^{n_k \times d} \rightarrow \mathbb{R}^{1 \times d'}\) or equivariant \(\mathbb{R}^{n_k \times d} \rightarrow \mathbb{R}^{n_k \times d'}\) layers can be formulated using a slight generalization of the previous analysis. The operator \(\mathbf{P} \ast \mathbf{A}\) is defined to act only on the nodal indices, i.e., \(i_1, \ldots, i_k\) (the first \(k\) indices). Explicitly, the \((p(i_1), p(i_2), \ldots, p(i_k), i_{k+1})\)-th entry of \(\mathbf{P} \ast \mathbf{A}\) equals the \((i_1, i_2, \ldots, i_k, i_{k+1})\)-th entry of \(\mathbf{A}\).

Invariance is now formulated exactly as before, equation 5 namely \(\mathbf{L}\mathbf{vec}(\mathbf{P} \ast \mathbf{A}) = \mathbf{Lvec}(\mathbf{A})\). The matrix that corresponds to \(\mathbf{P}\ast\) acting on \(\mathbb{R}^{n_k \times d}\) in the standard basis is \(\mathbf{P}^T \otimes \mathbf{I}_d\) and therefore \(\mathbf{L}\left(\mathbf{P}^T \otimes \mathbf{I}_d\right)\mathbf{vec}(\mathbf{A}) = \mathbf{Lvec}(\mathbf{A})\). Since this is true for all \(\mathbf{A}\) we have \(\left(\mathbf{P}^T \otimes \mathbf{I}_d\right)\mathbf{vec}(\mathbf{L}) = \mathbf{vec}(\mathbf{L})\), using the properties of the Kronecker product. Equivariance is written as in equation 6 \([\mathbf{Lvec}(\mathbf{P} \ast \mathbf{A})] = \mathbf{P} \ast [\mathbf{Lvec}(\mathbf{A})]\). In matrix form, the equivariance equation becomes \(\mathbf{L}\mathbf{vec}(\mathbf{P} \ast \mathbf{A}) = \mathbf{P} \ast \mathbf{Lvec}(\mathbf{A})\), since this is true for all \(\mathbf{A}\) and using the properties of the Kronecker product again we get to \(\mathbf{P}^T \otimes \mathbf{I}_d \otimes \mathbf{P} \otimes \mathbf{I}_{d'}\mathbf{vec}(\mathbf{L}) = \mathbf{vec}(\mathbf{L})\). The basis (with biases) to the solution space of these fixed-point equations is defined. We use \(\mathbf{a}, \mathbf{b} \in [n]^{k}, i, j \in [d], i', j' \in [d'], \lambda \in [n]^{k} / \sim, \mu \in [n]^{2k} / \sim\).

\[
\begin{align*}
\text{invariant:} \quad & \mathbf{B}_{\alpha,j,j'} = \begin{cases} 1 & \alpha \in \lambda, \ i=j, \\ 0 & \text{otherwise} \end{cases} \\
\text{equivariant:} \quad & \mathbf{B}_{\mu,j,j'} = \begin{cases} \frac{1}{\lambda} & (\alpha, \beta) \in \mu, \ i=j, \\ 0 & \text{otherwise} \end{cases} \\
& \mathbf{C}_{i,j,j'} = \begin{cases} 1 & i'=j' \\ 0 & \text{otherwise} \end{cases} \\
& \mathbf{C}_{b,i,j'} = \begin{cases} 1 & \beta \in \lambda, \ i'=j' \\ 0 & \text{otherwise} \end{cases}
\end{align*}
\]

(9a) (9b)

Note that these basis elements are similar to the ones in equation 8 with the difference that we have different basis tensor for each pair of input \(j\) and output \(j'\) feature channels.

An invariant (equation 10a) or equivariant (equation 10b) linear layer \(L\) including the biases can be written as follows for input \(\mathbf{A} \in \mathbb{R}^{n \times d}\):

\[
\begin{align*}
L(\mathbf{A})_{j'} &= \sum_{\alpha,i} \mathbf{T}_{\alpha,i} \mathbf{A}_{\alpha,i} + \mathbf{Y}; \\
\mathbf{T} &= \sum_{\lambda,j,j'} w_{\lambda,j,j'} \mathbf{B}_{\lambda,j,j'}; \\
\mathbf{Y} &= \sum_{j'} b_{j'} \mathbf{C}_{i,j,j'}; \\
L(\mathbf{A})_{b,j'} &= \sum_{\alpha,i} \mathbf{T}_{\alpha,i} \mathbf{A}_{\alpha,i} + \mathbf{Y}; \\
\mathbf{T} &= \sum_{\mu,j,j'} w_{\mu,j,j'} \mathbf{B}_{\mu,j,j'}; \\
\mathbf{Y} &= \sum_{j'} b_{j'} \mathbf{C}_{\mu,j,j'}.
\end{align*}
\]

(10a) (10b)

where the learnable parameters are \(w \in \mathbb{R}^{(k) \times d \times d'}\) and \(b \in \mathbb{R}^{d'}\) for a single linear invariant layer \(\mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{d'}\); and it is \(w \in \mathbb{R}^{(2k) \times d \times d'}\) and \(b \in \mathbb{R}^{(2k) \times d'}\) for a single linear equivariant layer \(\mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times d'}\). The natural generalization of theorem 1 to include bias and features is therefore:

**Theorem 2.** The space of invariance (equivariant) linear layers \(\mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{d'}\) is of dimension \(dd'v(k) + d'\) for invariance: \(dd'v(2k) + d'v(k)\)) with basis elements defined in equation 9, equation 10a, 10b show the general form of such layers.

Since, by similar arguments to proposition 2, the purely linear parts \(\mathbf{B}\) and biases \(\mathbf{C}\) in equation 9 are independent solutions to the relevant fixed-point equations, theorem 2 will be proved if their number equals the dimension of the solution space of these fixed-point equations, namely \(dd'v(k)\) for purely linear part and \(d'\) for bias in the invariant case, and \(dd'v(2k)\) for purely linear and \(d'v(k)\) for bias in the equivariant case. This can be shown by repeating the arguments of the proof of proposition 2 slightly adapted to this case, or by a combinatorial identity we show in Appendix 5.

For example, figure 1 depicts the 15 basis elements for linear equivariant layers \(\mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}\) taking as input edge-value (order-2) tensor data \(\mathbf{A} \in \mathbb{R}^{n \times n}\) and outputting the same dimension tensor. The basis for the purely linear part are shown as \(n^2 \times n^2\) matrices while the bias part as \(n \times n\) matrices (far right); the size of the node set is \(|\mathcal{V}| = n = 5\).

Mixed order equivariant layers. Another useful generalization of order equivariant linear layers is to linear layers between different order tensor layers, that is, \(L : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}\), where \(l \neq k\). For example, one can think of a layer mapping an adjacency matrix to per-node features. For simplicity we will discuss the purely linear scalar-valued case, however generalization to include bias and/or general feature vectors can be done as discussed above. Consider the matrix \(\mathbf{L} \in \mathbb{R}^{n \times n \times n_k}\) representing the linear layer \(L\), using the renumbering operator, \(\mathbf{P}\ast\), order equivariance is equivalent to \([\mathbf{Lvec}(\mathbf{P} \ast \mathbf{A})] = \mathbf{P} \ast [\mathbf{Lvec}(\mathbf{A})]\). Note that while this equation looks identical to equation 6 it is
nevertheless different in the sense that the $P_\star$ operator in the l.h.s. of this equation acts on $k$-order tensors while the one on the r.h.s. acts on $l$-order tensor. Still, we can transform this equation to a matrix equation as before by remembering that $P^{T \otimes k}$ is the matrix representation of the renumbering operator $P_\star$ acting on $k$-tensors in the standard basis. Therefore, repeating the arguments in proof of proposition 1 equivariance is equivalent to $P^\otimes (k+l) \text{vec}(L) = \text{vec}(L)$, for all permutation matrices $P$. This equation is solved as in section 3.1. The corresponding bases to such equivariant layers are computed as in equation 9b with the only difference that now $a \in [n]^k$, $b \in [n]^l$, and $\mu \in [n]^{k+l}/\sim$.

4 EXPERIMENTS

Implementation details. We implemented our method in Tensorflow (Abadi et al., 2016). The equivariant linear basis was implemented efficiently using basic row/column/diagonal summation operators, see appendix A for details. The networks we used are composition of $1 - 4$ equivariant linear layers with ReLU activation between them for the equivariant function setting. For invariant operators, see appendix A for details. The networks we used are composition of $\mu$.

We trained networks with $4$ non-symmetric part of the matrix. Indeed, Hartford et al. (2018) possess only $\mu$.

Hartford et al. (2018) seem to fail approximating rather simple functions that depend on the diagonal or the $\mu$.

Thus, although equivariant, Hartford et al. (2018) produce comparable errors to the trivial predictor. Thus, although equivariant, Hartford et al. (2018) seem to fail approximating rather simple functions that depend on the diagonal or the non-symmetric part of the matrix. Indeed, Hartford et al. (2018) possess only $4$ equivariant basis elements (with additional $1$ bias) in this case while we have $16$ basis elements (with $2$ biases).

Table 1: Comparison to baseline methods on synthetic experiments.

<table>
<thead>
<tr>
<th># Layers</th>
<th>Symmetric projection</th>
<th>Diagonal extraction</th>
<th>Max singular vector</th>
<th>Trace</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Trivial predictor</td>
<td>4.17</td>
<td>4.17</td>
<td>4.17</td>
<td>0.21</td>
</tr>
<tr>
<td>Hartford et al.</td>
<td>2.09</td>
<td>2.09</td>
<td>2.09</td>
<td>0.81</td>
</tr>
<tr>
<td>Ours</td>
<td>1E-05</td>
<td>7E-06</td>
<td>2E-05</td>
<td>8E-06</td>
</tr>
</tbody>
</table>

Synthetic datasets. We tested our method on several synthetic equivariant and invariant graph functions that highlight the differences in expressivity between our linear basis and the basis of Hartford et al. (2018). Given an input matrix data $A \in \mathbb{R}^{n \times n}$ we considered: (i) projection onto the symmetric matrices $\frac{1}{2}(A + A^T)$; (ii) diagonal extraction $\text{diag}(\text{diag}(A))$ (keeps only the diagonal and plugs zeros elsewhere); (iii) computing the maximal right singular vector $\arg \max_{\|v\|_2=1} \|Av\|_2$; and (iv) computing the trace $\text{tr}(A)$. Tasks (i)-(iii) are equivariant while task (iv) is invariant. We created accordingly $4$ datasets with $10K$ train and $1K$ test examples of $40 \times 40$ matrices; for tasks (i), (ii), (iv) we used i.i.d. random matrices with uniform distribution in $[0, 10]$; we used mean-squared error (MSE) as loss; for task (iii) we random matrices with uniform distribution of singular values in $[0, 0.5]$ and spectral gap $\geq 0.5$; due to sign ambiguity in this task we used cosine loss of the form $l(x, y) = 1 - \langle x, y \rangle / \|x\|_2 \|y\|_2^2$.

We trained networks with $1$, $2$, and $3$ hidden layers with $8$ feature channels each and a single fully-connected. Both ours and Hartford et al. (2018) use the same architecture but with different bases for the linear layers. Table 1 logs the best mean-square error of each method over a set of hyperparameters. We add the MSE for the trivial mean predictor. Using the bases of Hartford et al. (2018) produce comparable errors to the trivial predictor. Thus, although equivariant, Hartford et al. (2018) seem to fail approximating rather simple functions that depend on the diagonal or the non-symmetric part of the matrix. Indeed, Hartford et al. (2018) possess only $4$ equivariant basis elements (with additional $1$ bias) in this case while we have $16$ basis elements (with $2$ biases).

Table 2: Generalization.

<table>
<thead>
<tr>
<th>n</th>
<th>sym</th>
<th>svd</th>
<th>diag</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>3.8E-05</td>
<td>0.0013</td>
<td>0.0150</td>
</tr>
<tr>
<td>40</td>
<td>7E-06</td>
<td>0.0084</td>
<td>1.5E-05</td>
</tr>
<tr>
<td>50</td>
<td>0.0055</td>
<td>0.0096</td>
<td>0.0055</td>
</tr>
</tbody>
</table>

Graph classification. We tested our method on standard benchmarks of graph classification. We use $8$ different real world datasets from the benchmark of Yanardag & Vishwanathan (2015): five of these datasets originate from bioinformatics while the other three come from social networks. In all datasets the adjacency matrix of each graph is used as input and a categorical label is assigned as output. In the bioinformatics datasets node labels are also provided as inputs. These node labels can be used in our framework by placing their 1-hot representations on the diagonal of the input.
Under review as a conference paper at ICLR 2019

Lastly, we provide a generalization of our framework to data that is given on tuples of nodes (Shervashidze et al., 2009), Random Walk Kernel (RW) (Vishwanathan et al., 2010), Propagation (Niepert et al., 2016), DCNN (Atwood & Towsley, 2016), ECC (Simonovsky & Komodakis, 2017), We compared our results to seven deep learning methods: DGCNN (Zhang et al., 2018), PSCN (Niepert et al. [2016], DCNN [Atwood & Towsley [2016], ECC [Simonovsky & Komodakis [2017], DGK [Yanardag & Vishwanathan [2015], DiffPool [Ying et al. [2018] and CCN [Kondor et al. [2018]. We also compare our results to four popular graph kernel methods: Graphlet Kernel (GK) (Shervashidze et al. [2009], Random Walk Kernel (RW) (Vishwanathan et al. [2010], Propagation Kernel (PK) (Neumann et al. [2016], and Weisfeiler-lehman kernels (WL) (Shervashidze et al. [2011]. Our method achieved results superior to the state-of-the-art on the three social networks datasets, and slightly worse results than state-of-the-art on the biological datasets.

5 Generalizations to multi-node sets

Lastly, we provide a generalization to our framework to data that is given on tuples of nodes from a collection of node sets \( \mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m \) of sizes \( n_1, n_2, \ldots, n_m \) (resp.), namely \( \mathbf{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m} \). We characterize invariant linear layers \( \mathbf{L} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m} \rightarrow \mathbb{R} \) and equivariant linear layer \( \mathbf{L} : \mathbb{R}^{n_1^1 \times n_2^1 \times \cdots \times n_m^1} \rightarrow \mathbb{R}^{n_1^2 \times n_2^2 \times \cdots \times n_m^2} \), where for simplicity we do not discuss features that can be readily added as discussed in section 3. Note that the case of \( k_i = l_i = 1 \) for all \( i = 1, \ldots, m \) is treated in [Hartford et al. [2018]). The reordering operator now is built out of permutation matrices \( \mathbf{P} \in \mathbb{R}^{n_1 \times n_m} \), \( \mathbf{P}_i \) denotes the permutation), \( i = 1, \ldots, m \), denoted \( \mathbf{P}_{1 \cdots m} \), and defined as follows: the \( (p_1(a_1), p_2(a_2), \ldots, p_m(a_m)) \)-th entry of the tensor \( \mathbf{P}_{1 \cdots m} \ast \mathbf{A} \), where \( a_i \in [n_i][k_i] \), is defined to be the \( (a_1, a_2, \ldots, a_m) \)-th entry of the tensor \( \mathbf{A} \). Rewriting the invariant and equivariant equations, i.e., equation 2 in matrix format, similarly to before, we get the fixed-point equations: \( \mathbf{M} \vec{v}(\mathbf{L}) = \vec{v}(\mathbf{L}) \) for invariant, and \( \mathbf{M} \otimes \mathbf{M} \vec{v}(\mathbf{L}) = \vec{v}(\mathbf{L}) \) for equivariant, where \( \mathbf{M} = \mathbf{P} \otimes k_1 \otimes \cdots \otimes \mathbf{P}_{m} \otimes k_m \). The solution of these equations would be linear combinations of basis tensors similar to equation 2 of the form

\[
\text{invariant: } \mathbf{B}_{1, \ldots, \lambda_m} = \begin{cases} 
1 & a_i \in \lambda_i, \forall i \in [1, \ldots, \mu] \\
0 & \text{otherwise}
\end{cases} \quad \text{equivariant: } \mathbf{B}_{a_1, \ldots, a_m, b_1, \ldots, b_m} = \begin{cases} 
1 & (a_i, b_i) \in \mu_i, \forall i \\
0 & \text{otherwise}
\end{cases}
\]

where \( \lambda \in [n_1][k_1], \mu_i \in [n_i][k_i] \), \( a_i \in [n_i][k_i] \), \( b_i \in [n_i][l_i] \). The number of these tensors is \( \prod_{i=1}^{m} b(i) \) for invariant layers and \( \prod_{i=1}^{m} b(i) \cdot l(i) \) for equivariant layers. Since these are all linear independent (pairwise disjoint support of non-zero entries) we need to show that their number equal the dimension of the solution of the relevant fixed-point equations above. This can be done again by similar arguments to the proof of proposition 2 or as shown in appendix B. To summarize:

**Theorem 3.** The linear space of invariant linear layers \( \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m} \rightarrow \mathbb{R} \) is of dimension \( \prod_{i=1}^{m} b(i) \). The equivariant linear layers \( \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m} \rightarrow \mathbb{R}^{n_1^2 \times n_2^2 \times \cdots \times n_m^2} \) has dimension \( \prod_{i=1}^{m} b(i) \cdot l(i) \). Orthogonal bases for these layers are listed in equation 7.
REFERENCES


**APPENDIX A EFFICIENT IMPLEMENTATION OF LAYERS**

For fast execution of order-2 layers we implemented the following 15 operations which can be easily shown to span the basis discussed in the paper. We denote by \( \mathbf{1} \in \mathbb{R}^n \) the vector of all ones.

1. The identity and transpose operations: 
   \[ L(A) = A, \quad L(A) = A^T. \]
2. The diag operation: 
   \[ L(A) = \text{diag}(\text{diag}(A)). \]
3. Sum of rows replicated on rows/ columns/ diagonal: 
   \[ L(A) = A\mathbf{1}\mathbf{1}^T, \quad L(A) = (A\mathbf{1})^T, \quad L(A) = \text{diag}(A\mathbf{1}). \]
4. Sum of columns replicated on rows/ columns/ diagonal: 
   \[ L(A) = A^T\mathbf{1}\mathbf{1}^T, \quad L(A) = (A^T\mathbf{1})^T, \quad L(A) = \text{diag}(A^T\mathbf{1}). \]
5. Sum of all elements replicated on all matrix/ diagonal: 
   \[ L(A) = (1^T\mathbf{1})\cdot \mathbf{1}\mathbf{1}^T, \quad L(A) = (1^T\mathbf{1})\cdot \text{diag}(1). \]
6. Sum of diagonal elements replicated on all matrix/diagonal: 
   \[ L(A) = (1^T\text{diag}(A))\cdot \mathbf{1}\mathbf{1}^T, \quad L(A) = (1^T\text{diag}(A))\cdot \text{diag}(1). \]
7. Replicate diagonal elements on rows/columns: 
   \[ L(A) = \text{diag}(A)\mathbf{1}^T, \quad L(A) = \mathbf{1}\text{diag}(A)^T. \]

We normalize each operation to have unit max operator norm. We note that in case the input matrix is symmetric, our basis reduces to 11 elements in the first layer. If we further assume the matrix has zero diagonal we get a 6 element basis in the first layer. In both cases our model is more expressive than the 4 element basis of Hartford et al. (2018) and as the output of the first layer (or other inner states) need not be symmetric nor have zero diagonal the deeper layers can potentially make good use of the full 15 element basis.

**APPENDIX B INVARIANT AND EQUIVARIANT SUBSPACE DIMENSIONS**

We prove a useful combinatorial fact as a corollary of proposition 2. This fact will be used later to easily compute the dimensions of more general spaces of invariant and equivariant linear layers. We use the fact that if \( V \) is a representation of a finite group \( G \) then 
\[
\phi = \frac{1}{|G|} \sum_{g \in G} g \in \text{End}(V)
\]
(12)
is a projection onto \( V^G = \{v \in V \mid gv = v, \forall g \in G\} \), the subspace of fixed points in \( V \) under the action of \( G \), and consequently that \( \text{tr}(\phi) = \dim(V^G) \) (see Fulton & Harris 2013 for simple proofs).

**Proposition 3.** The following formula holds:
\[
\frac{1}{n!} \sum_{P \in \Pi_n} \text{tr}(P)^k = b(k),
\]
where \( \Pi_n \) is the matrix permutation group of dimensions \( n \times n \).
Proof. In our case, the vector space is the space of order-$k$ tensors and the group acting on it is the matrix group $G = \{ P^{\otimes k} | P \in \Pi_m \}$.

$$\dim(V^G) = \text{tr}(\phi) = \frac{1}{|G|} \sum_{g \in G} \text{tr}(g) = \frac{1}{n!} \sum_{P \in \Pi_n} \text{tr}(P^{\otimes k}) = \frac{1}{n!} \sum_{P \in \Pi_n} \text{tr}(P)^k;$$

where we used the multiplicative law of the trace with respect to Kronecker product. Now we use proposition\textsuperscript{2} noting that in this case $V^G$ is the solution space of the fixed-point equations. Therefore, $\dim(V^G) = b(k)$ and the proof is finished.

Recall that for a permutation matrix $P$, $\text{tr}(P) = |\{ i \in [n] \text{ s.t. } P \text{ fixes } e_i \}|$. Using this, we can interpret the equation in proposition\textsuperscript{3} as the $k$-th moment of a random variable counting the number of fixed points of a permutation, with uniform distribution over the permutation group. Proposition\textsuperscript{3} proves that the $k$-th moment of this random variable is the $k$-th Bell number.

We can now use proposition\textsuperscript{3} to calculate the dimensions of two linear layer spaces: (i) Equivariant layers acting on order-$k$ tensors with features (as in \textsuperscript{3}); and (ii) multi-node sets (as in section 5).

**Theorem 2.** The space of invariant (equivariant) linear layers $\mathbb{R}^{nk,d} \rightarrow \mathbb{R}^{d'}$ ($\mathbb{R}^{nk,d} \rightarrow \mathbb{R}^{nk,d'}$) is of dimension $dd' b(k)$ (for equivariant: $dd' b(2k) + d'b(k)$) with basis elements defined in equation \textsuperscript{9}; equations \textsuperscript{10a}\textsuperscript{10b} show the general form of such layers.

Proof. We prove the dimension formulas for the invariant case. The equivariant case is proved similarly. The solution space for the fixed point equations is the set $V^G$ for the matrix group $G = \{ P^{\otimes k} \otimes I_d \otimes I_{d'} | P \in \Pi_n \}$. Using the projection formula\textsuperscript{12} we get that the dimension of the solution subspace, which is the space of invariant linear layers, can be computed as follows:

$$\dim(V^G) = \frac{1}{n!} \sum_{P \in \Pi_n} \text{tr}(P)^k \text{tr}(I_d) \text{tr}(I_{d'}) = \left( \frac{1}{n!} \sum_{P \in \Pi_n} \text{tr}(P)^k \right) \text{tr}(I_d) \text{tr}(I_{d'}) = d \cdot d' \cdot b(k).$$

**Theorem 3.** The linear space of invariant linear layers $L : \mathbb{R}^{n_{k1} \times n_{k2} \times \cdots \times n_{km}} \rightarrow \mathbb{R}$ is of dimension $\prod_{i=1}^{m} b(k_{i})$. The equivariant linear layers $L : \mathbb{R}^{n_{k1} \times n_{k2} \times \cdots \times n_{km}} \rightarrow \mathbb{R}^{n_{l1} \times n_{l2} \times \cdots \times n_{lm}}$ has dimension $\prod_{i=1}^{m} b(k_{i} + l_{i})$. Orthogonal bases for these layers are listed in equation \textsuperscript{11}.

Proof. In this case we get the fixed-point equations: $M \text{vec}(L) = \text{vec}(L)$ for invariant, and $M \otimes M \text{vec}(L) = \text{vec}(L)$ for equivariant, where $M = P_{1}^{\otimes k_{1}} \otimes \cdots \otimes P_{m}^{\otimes k_{m}}$. Similarly to the previous theorem, plugging $M$ into equation\textsuperscript{12} using the trace multiplication rule and proposition\textsuperscript{3} we get the above formulas.