A Graph Neural Networks (GNNs)

We consider standard message-passing graph neural networks (MPNNs) [19–21] defined as follows. A L-layer MPNN maps input $X \in \mathbb{R}^{N\times d}$ to output $Y \in \mathbb{R}^{N\times k}$ following an iterative scheme: At initialization, $\mathbf{h}^{(0)} = X$; At each iteration l, the embedding for node i is updated to

$$\mathbf{h}_i^{(l)} = \phi\left(\mathbf{h}_i^{(l-1)}, \sum_{j\in\mathcal{N}(i)} \psi\left(\mathbf{h}_i^{(l-1)}, \mathbf{h}_j^{(l-1)}, A_{[i,j]}\right)\right), \tag{7}$$

where ϕ, ψ are the update and message functions, $\mathcal{N}(i)$ denotes the neighbors of node i, and $A_{[i,j]}$ represents the (i,j)-edge weight. MPNNs typically have two key design features: (1) ϕ, ψ are *shared* across all nodes in the graph, typically chosen to be a linear transformation or a multi-layer perceptions (MLPs), known as *global weight sharing*; (2) the graph A is used for (spatial) convolution.

B Parameterization of Linear Equivariant Maps

We consider a group $\mathcal{G}$ acting on spaces $\mathcal{X}$ and $\mathcal{Y}$ via representations ϕ and ψ, respectively. Our goal is to find the linear equivariant maps $f : \mathcal{X} \to \mathcal{Y}$ such that $f(\phi(g)x) = \psi(g)f(x)$ for all $g \in \mathcal{G}$ and $x \in \mathcal{X}$. The standard way to do this, used extensively in the equivariant machine learning literature (e.g. [40, 43]), is to decompose ϕ and ψ in irreducibles and use Schur's lemma.

In a nutshell, a group representation φ is an homomorphism $\mathcal{G} \to \mathrm{GL}(V)$ (sometimes mathematicians say that V is a representation of $\mathcal{G}$, but we need to know the homomorphism φ too). One way to interpret the group homomorphism (i.e. $\varphi(gh) = \varphi(g) \circ \varphi(h)$) is that the group multiplication corresponds to the composition of linear invertible maps (i.e. matrix multiplication). A linear subspace W of V is said to be a subrepresentation of φ if $\varphi(\mathcal{G})(W) \subset W$. A irreducible representation is one that only has itself and the trivial subspace as subrepresentations.

Schur's lemma states that if V, W are vector spaces over $\mathbb{C}$ and φ_V, φ_W are irreducible representations, then either (1) φ_V and φ_W are not isomorphic as representations (and the only linear equivariant map between V, W is the zero map), or (2) φ_V and φ_W are isomorphic and the only non-trivial equivariant maps are of the form $\lambda\, I$ where $\lambda \in \mathbb{C}$ and I is the identity (See Chapter 1 of [60]).

Now given $\mathcal{G}$ acting on spaces $\mathcal{X}$ and $\mathcal{Y}$ via representations ϕ and ψ, respectively. Then one can decompose ϕ and ψ in irreducibles over $\mathbb{C}$

$$\phi = \oplus_{k=1}^{\ell} a_k \mathcal{T}_k \quad \psi = \oplus_{k=1}^{\ell} b_k \mathcal{T}_k$$

(this notation assumes the same irreducibles appear in both decompositions, which can be done if we allow some of the a_k and b_k to be zero). And then one can parameterize the equivariant maps by having one complex parameter per irreducible that appears in both decompositions. These ideas can be applied to real spaces.

Then finding the linear equivariant maps reduces to decomposing the corresponding representations in irreducibles. In the next sections we explain in detail how to do this for the specific problems described in this paper. The appendix is organized as follows: We first show how to parameterize equivariant linear layers for Abelian group (Section B.1.1), and then provide the end-to-end design of equivariant graph networks $\mathcal{G}$-Net (Section B.3).

B.1 Equivariant Linear Maps via Isotypical Decomposition

In this section, we assume that the graph adjacency matrix A has distinct eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. Then $\mathcal{A}_G$ is an Abelian group (Lemma 3.8.1, notes). Under this assumption, we present the construction of approximately equivariant graph networks using isotypical decomposition (i.e. decomposition into isomorphism classes of irreducible representations) and group characters. We remark that such construction extends to non-Abelian groups and refer the interested reader to [68], but we omit it here for the ease of exposition.

B.1.1 Equivariant Linear Layers for Abelian Group

We consider the simplest setting where $f : \mathbb{R}^N \to \mathbb{R}^N$ is a linear function that maps signals on the node level. Let $x \in \mathbb{R}^N$ be the node features, then equivariance requires

$$f(g\,x) = g\,f(x) \quad \text{for all } g \in \mathcal{A}_G. \tag{8}$$

To construct linear equivariant functions f, our roadmap is outlined as follows:

1. Decompose the vector space $\mathcal{X} = \mathbb{R}^N$ into a sum of components such that different components cannot be mapped to each other equivariantly (also known as the isotypic decomposition);
2. Given $\mathcal{X} = \oplus_i \mathcal{X}_i$ an isotypic representation, we then parameterize f by linear maps at each $\mathcal{X}_i$ such that for all i, $f(\mathcal{X}_i) \subseteq \mathcal{X}_i$.

To this end, we need the following definitions.

Definition 5. *($\mathcal{G}$-module, [68, Defn 1.3.1]) Let $\mathcal{X}$ be a vector space and $\mathcal{G}$ be a group. We say the vector space $\mathcal{X}$ is a $\mathcal{G}$-module or $\mathcal{X}$ carries a representation of $\mathcal{G}$ if there is a group homomorphism $\rho : \mathcal{G} \to GL(\mathcal{X})$, where GL denotes the General Linear group. Equivalently, if the following holds:*

1. *$gv \in \mathcal{X}$,*
2. *$g(cv + dw) = c(gv) + d(gw)$,*
3. *$(gh)v = g(hv)$,*
4. *$ev = v$*

for all $g, h \in \mathcal{G}; v, w \in \mathcal{X}$ and scalars $c, d \in \mathbb{C}$ ($e \in \mathcal{G}$ denotes the identity element).

In what follows, we consider $\mathcal{X} = \mathbb{R}^N$ carries a representation of G.

Definition 6. *(Group characters) Let $X(g), g \in \mathcal{G}$ be a matrix representation of a group element. Then the character of X is $\chi(g) \coloneqq \operatorname{tr} X(g)$.*

Definition 7. *(Group orbits) Let $\mathcal{X}$ be a vector space and $\mathcal{G}$ be a group. The group orbit of an element $x \in \mathcal{X}$ is $O(x) \coloneqq \{gx : g \in \mathcal{G}\}$.*

Let $g_1, \ldots, g_s$ be the generators of $\mathcal{A}_G \subset (\mathcal{S}_2)^n$, or simply $\mathcal{A}_G \equiv (\mathcal{S}_2)^k$ for some $k \le n$. Since $\mathcal{A}_G$ is abelian, any irreducible representation is 1-dimensional [60, p.8]. In other words, the irreducible representations of an abelian group are homomorphisms

$$\rho : \mathcal{A}_G \to \mathbb{C}. \tag{9}$$

Since all the elements of the group $\mathcal{A}_G = (\mathcal{S}_2)^k$ is of order 1 or 2, the homomorphisms are $\rho : \mathcal{A}_G \to \{\pm 1\} \subset \mathbb{R}$. By Defn 6, the irreducible characters (i.e., characters of irreducible matrix representation) are also homomorphisms $\rho : \mathcal{A}_G \to \{\pm 1\}$. In other words, $\chi(g) \in \{\pm 1\}$for all $g \in \mathcal{A}_G$. Then we can write down the $2^k \times 2^k$ character table, where the rows are the characters χ, and the columns are the group elements $g \in \mathcal{A}_G$ (see Table 3 as an example). Now, define the projection onto the isotypic component of the representation X as

$$P_\chi \coloneqq \frac{\deg(X)}{|\mathcal{A}_G|} \sum_{g \in \mathcal{A}_G} \overline{\chi(g)}\, g = \frac{1}{|\mathcal{A}_G|} \sum_{g \in \mathcal{A}_G} \chi(g)\, g, \tag{10}$$

where the second equality uses the fact that $\mathcal{A}_G$ is abelian.

Intuitively, applying P_χ on $\mathcal{X} = \operatorname{span}(\{e_1, \ldots, e_N\})$ picks out all $v \in \mathcal{X}$ that stays in the same subspace defined by the group character χ. (Note that for the $(\mathcal{S}_2)^k$ case $\chi^{-1}(g) = \chi(g)$ since $\chi(g) \in \{\pm 1\}$).

We are ready to present the precise construction of linear equivariant map f with respect to an Abelian group:

Lemma 5. f is linear, equivariant with respect to the abelian group $\mathcal{A}_G$ if and only if f can be written as (12) in Algorithm 1.

Algorithm 1 Parameterizing linear equivariant functions $f : \mathbb{R}^N \to \mathbb{R}^N$ for abelian group

Require: Abelian group $\mathcal{A}_G = (\mathcal{S}_2)^k$

1. Construct the character table of χ_{irreps} for $\mathcal{A}_G$, i.e. $\chi_i : \mathcal{A}_G \to \{\pm 1\}$ $i = 1, \dots \ell$;
2. For each character χ_i in the character table, compute the projection matrix
$$P_{\chi_i}(\mathcal{X}) = [P_{\chi_i}(e_1); \dots; P_{\chi_i}(e_N)] \in \mathbb{R}^{N \times N}; \tag{11}$$
followed by computing the basis from $P_{\chi_i}(\mathcal{X})$ and call it $\mathcal{X}_{\chi_i} = [b^{(1)}_{\chi_i}, \dots, b^{(K_i)}_{\chi_i}]$.
3. $\mathcal{X} = \oplus_{i=1}^{\ell} \mathcal{X}_{\chi_i}$ where $\mathcal{X}_{\chi_i}$ are the isotypic component. Then f is any linear function satisfying that $f(\mathcal{X}_{\chi_i}) \subseteq \mathcal{X}_{\chi_i}$ for all $i = 1, \dots, \ell$. In particular, in the basis $[b^{(s)}_{\chi_i}]_{1 \leq i \leq \ell, 1 \leq s \leq K_i}$ f can be written as a block diagonal matrix $\mathbb{R}^{n \times n}$ with each block M_{χ_i} being the linear map from $\mathcal{X}_{\chi_i} \to \mathcal{X}_{\chi_i}$,
$$f = \begin{bmatrix} M_{\chi_1} & & & \\ & M_{\chi_2} & & \\ & & \ddots & \\ & & & M_{\chi_\ell} \end{bmatrix} \tag{12}$$

return f

	e	σ
χ_e	1	1
χ_2	1	-1

Table 3: Character table for $\text{aut}(P_4) \cong Z_2$

Proof. By construction in Algorithm 1, f is linear and equivariant. To show the converse, since $\mathcal{A}_G$ is abelian with all irreducible representations being one-dimensional, for $\mathcal{X}_{\chi_1} \ncong \mathcal{X}_{\chi_2}$, we have

$$g\, v_1 = \lambda_1(g)\, v_1, \quad \text{for all } g \in \mathcal{G}, v_1 \in \mathcal{X}_{\chi_1}, \tag{13}$$
$$g\, v_2 = \lambda_2(g)\, v_2, \quad \text{for all } g \in \mathcal{G}, v_2 \in \mathcal{X}_{\chi_2}, \tag{14}$$

where there exists some $g \in \mathcal{G}$ such that $\lambda_1(g) \neq \lambda_2(g)$. To show f being linear and equivariant implies for all $v \in \mathcal{X}_\chi$, $f(v) \in \mathcal{X}_\chi$, we prove by contradiction. Without loss of generality, suppose

$$f(v_{\chi_1}) = \alpha_1 v_{\chi_1} + \alpha_2 v_{\chi_2}, \tag{15}$$

for some scalars α_1, α_2 and $v_{\chi_1} \in \mathcal{X}_{\chi_1}, v_{\chi_2} \in \mathcal{X}_{\chi_2}$. Then by (13), for all $g \in \mathcal{G}$,

$$f(g\, v_{\chi_1}) = f(\lambda_1(g)\, v_{\chi_1}) = \lambda_1(g) f(v_{\chi_1}) = \lambda_1(g)\alpha_1 v_{\chi_1} + \lambda_1(g)\alpha_2 v_{\chi_2}. \tag{16}$$

Now, since f is equivariant, for all $g \in \mathcal{G}$,

$$f(g\, v_{\chi_1}) = g f(v_{\chi_1}) = g(\alpha_1 v_{\chi_1} + \alpha_2 v_{\chi_2}) = \lambda_1(g)\alpha_1 v_{\chi_1} + \lambda_2(g)\alpha_2 v_{\chi_2}. \tag{17}$$

But there exists some $g' \in \mathcal{G}$ such that $\lambda_1(g') \neq \lambda_2(g')$, which leads to $f(g' v_{\chi_1}) \neq f(g' v_{\chi_1})$, a contradiction. One can easily extend the proof strategy to the general case for $f(v_{\chi_1}) = \sum_{i=1}^{l} v_{\chi_i}$. $\square$

Example B.1. Consider the path graph on 4 nodes (i.e., P_4). We have $\text{aut}(P_4) = \{e, (14)(23)\} \cong Z_2$.

Steps 1: Note that Z_2 is Abelian and thus all irreducible characters $\chi(g) \in \{\pm 1\}$, for all $g \in Z_2$. The character table is shown in Table 3.

Step 2: using (10) we have

$$P_{\chi_e}[e_1; e_2; e_3; e_4] = \frac{1}{2}\begin{bmatrix}1 & 0 & 0 & 1\\ 0 & 1 & 1 & 0\\ 0 & 1 & 1 & 0\\ 1 & 0 & 0 & 1\end{bmatrix} \text{ which yields basis } \mathcal{B}(P_{\chi_e}) = [e_1 + e_4; e_2 + e_3].$$

$$P_{\chi_2}[e_1; e_2; e_3; e_4] = \frac{1}{2}\begin{bmatrix}1 & 0 & 0 & -1\\ 0 & 1 & -1 & 0\\ 0 & -1 & 1 & 0\\ -1 & 0 & 0 & 1\end{bmatrix} \text{ which yields basis } \mathcal{B}(P_{\chi_2}) = [e_1 - e_4; e_2 - e_3].$$

Step 3: Parameterize $f : \mathbb{R}^4 \to \mathbb{R}^4$ by $f : \mathcal{B}(P_{\chi_e}) \to \mathcal{B}(P_{\chi_e})$ and $f : \mathcal{B}(P_{\chi_2}) \to \mathcal{B}(P_{\chi_2})$, i.e. for all $v \in \mathbb{R}^4$,let $v = c_1(e_1 + e_4) + \ldots + c_4(e_2 - e_3)$, then

$$f(v) = \begin{bmatrix}\alpha_1 & \alpha_2\\ \alpha_3 & \alpha_4\end{bmatrix}[c_1; c_2] + \begin{bmatrix}\alpha_5 & \alpha_6\\ \alpha_7 & \alpha_8\end{bmatrix}[c_3; c_4], \tag{18}$$

where $\alpha_1, \ldots, \alpha_8$ are (learnable) real scalars. Now f is linear, equivariant by construction.

B.2 Equivariant Linear Map for Symmetries Induced by Graph Coarsening

In this section, we present the construction of equivariant linear maps for some examples using the symmetry group induced by graph coarsening (Defn 3). Recall the symmetry group with M clusters of G (with the associated coarsened graph G') is given by

$$\mathcal{G}_{G\to G'} := \Big(\mathcal{S}_1 \times \mathcal{S}_2 \ldots \times \mathcal{S}_M\Big) \rtimes \overline{\mathcal{A}}_{G'} \subset \mathcal{S}_N.$$

Here we assume that $\overline{\mathcal{A}}_{G'}$ is trivial and we show how to parameterize equivariant functions with respect to products of permutations. In more general cases, for instance if $\overline{\mathcal{A}}_{G'}$ is abelian, we can use a construction by Serre ([69] Section 8.2). For the ease of exposition, consider $X \in \mathbb{R}^N, Y \in \mathbb{R}^N$. Then any permutation-equivariant linear function $f : \mathbb{R}^N \to \mathbb{R}^N$ with respect to $\mathcal{G}_{G\to G'}$ admits the following block-matrix form:

$$f = \begin{bmatrix}f_{11} & f_{12} & \cdots & f_{1M}\\ f_{21} & f_{22} & \cdots & f_{2M}\\ & & \ddots & \\ f_{M1} & f_{M2} & \cdots & f_{MM}\end{bmatrix},\ f_{kk} = a_k\mathbf{I} + b_k\mathbb{1}\mathbb{1}^\top,\ f_{kl} = c_{kl}\mathbb{1}\mathbb{1}^\top \text{ for } k \neq l, \tag{19}$$

where f_{kl} are block matrices, and a_k, b_k, c_{kl} are scalars where $c_{kl} = c_{lk}$ if and only if the coarsened nodes $k, l \in G'$ are in the same group orbit. Figure 2 illustrates the block structure of f. This is due to (1) f_{kk} is a linear permutation-equivariant function if and only if its diagonal elements are the same and its off-diagonal elements are the same ([34, Lemma 3.]); (2) f_{kl} for $k \neq l$ is a constant matrix since nodes within a cluster are indistinguishable, and c_{kl} needs to satisfy the symmetry of $\overline{\mathcal{A}}_{G'}$.

Finally, we illustrate the linear equivariant layer for two-cluster graph coarsening. Without loss of generality, assume that the adjacency matrix A and the node signals X are ordered according to the cluster assignment (e.g., $X_{[1:|V_1|]}$ are node features for the first cluster, etc). Let $X_{(1)}, X_{(2)}$ denote the node features for the first and second cluster, $W^s_{(1)}, W^s_{(2)}$ denote the weights on the block diagonal for self-feature transformation, $W^n_{(1)}, W^n_{(2)}$ denote the weights on the block diagonal for within-cluster neighbors, and $W^n_{(12)}, W^n_{(21)}$ denote the weights off the block diagonal for across-cluster neighbors. Let I denote the identity matrix, and $\mathbf{1}_{(1)}, \mathbf{1}_{(2)}$ denote the all-ones matrices with the same size as the corresponding cluster. Recall $\odot$ denotes the element-wise multiplication of two matrices. Then the linear equivariant layer is parameterized as

$$A\odot I\begin{bmatrix}X_{(1)}W^s_{(1)}\\ X_{(2)}W^s_{(2)}\end{bmatrix} + A\odot\Big(\begin{bmatrix}\mathbf{1}_{(1)} & 0\\ 0 & \mathbf{1}_{(2)}\end{bmatrix} - I\Big)\begin{bmatrix}X_{(1)}W^n_{(1)}\\ X_{(2)}W^n_{(2)}\end{bmatrix} + A\odot\begin{bmatrix}0 & \mathbf{1}_{(2)}\\ \mathbf{1}_{(1)} & 0\end{bmatrix}\begin{bmatrix}X_{(1)}W^n_{(12)}\\ X_{(2)}W^n_{(21)}\end{bmatrix}. \tag{20}$$

B.3 Equivariant Layer for Human Skeleton Graph

We now apply the constructions above to our human skeleton graph described in Section 5.1. We first show how to parameterize all linear $\mathcal{A}_G$-equivariant functions. Observe that $\mathcal{A}_G \cong (\mathcal{S}_2)^2 =$

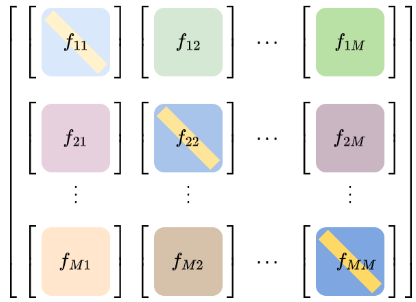


Figure 2: The block structure of linear equivariant function $f : \mathbb{R}^n \to \mathbb{R}^n$ with respect to $\mathcal{G}_{G \to G'}$ (where G, G' are asymmetric): Each diagonal block f_{kk} is diagonally constant and off-diagonally constant; Each off-diagonal block f_{kl} is a constant matrix.

$\{e, a, l, al\}$, where the nontrivial actions correspond to the **a**rm flip with respect to the spine, the **l**eg flip with respect to the spine, and their composition. To fix ideas, we first treat both input and output graph signals as vectors, and construct $\mathcal{A}_G$-equivariant linear maps $f : \mathbb{R}^{16} \to \mathbb{R}^{16}$.

Step 1: Obtain the character table for $(\mathcal{S}_2)^2$

	e	a	l	al
χ_e	1	1	1	1
χ_2	1	1	-1	-1
χ_3	1	-1	1	-1
χ_4	1	-1	-1	1

Table 4: Character table for $(\mathcal{S}_2)^2$

Step 2: Construct the basis for isotypic decomposition. Here we choose to index the leg joint pairs as $(1, 4), (2, 5), (3, 6)$, arm joint pairs as $(10, 13), (11, 14), (12, 15)$, and spline joints $0, 7, 8, 9$.

$$
\begin{aligned}
B &= [\mathcal{B}(P_{\chi_e}); \mathcal{B}(P_{\chi_2}); \mathcal{B}(P_{\chi_3}); \mathcal{B}(P_{\chi_4})] \text{ where} \\
\mathcal{B}(P_{\chi_e}) &= [(e_1 + e_4)/\sqrt{2}; \ldots; (e_{12} + e_{15})/\sqrt{2}; e_0; e_7; e_8; e_9] \in \mathbb{R}^{16 \times 10}. \\
\mathcal{B}(P_{\chi_2}) &= [(e_1 - e_4)/\sqrt{2}; (e_2 - e_5)/\sqrt{2}; (e_3 - e_6)/\sqrt{2}] \in \mathbb{R}^{16 \times 3}; \\
\mathcal{B}(P_{\chi_3}) &= [(e_{10} - e_{13})/\sqrt{2}; (e_{11} - e_{14})/\sqrt{2}; (e_{12} - e_{15})/\sqrt{2}] \in \mathbb{R}^{16 \times 3}; \\
\mathcal{B}(P_{\chi_4}) &= \emptyset
\end{aligned} \tag{21}
$$

Step 3: Parameterize $f : \mathbb{R}^{16} \to \mathbb{R}^{16}$ by $f : \mathcal{B}(P_{\chi_e}) \to \mathcal{B}(P_{\chi_e})$ and $f : \mathcal{B}(P_{\chi_2}) \to \mathcal{B}(P_{\chi_2})$, i.e. for all $v \in \mathbb{R}^{16}$, let $v = \mathcal{B}(P_{\chi_e})\, \boldsymbol{c_e} + \mathcal{B}(P_{\chi_2})\, \boldsymbol{c_2} + \mathcal{B}(P_{\chi_3})\, \boldsymbol{c_3}$, then

$$
f(v) = W_e\, \boldsymbol{c_e} + W_2\, \boldsymbol{c_2} + W_3\, \boldsymbol{c_3}, \tag{22}
$$

where $W_e \in \mathbb{R}^{10 \times 10}, W_2 \in \mathbb{R}^{3 \times 3}, W_3 \in \mathbb{R}^{3 \times 3}$ are (learnable) weight matrices. Now f expresses all linear, equivariant maps w.r.t $(\mathcal{S}_2)^2$.

The following calculation based on $f : \mathbb{R}^{16} \to \mathbb{R}^{16}$ shows how much degree of freedom (measured by learnable parameters) is gained by relaxing the symmetry from global (group $\mathcal{S}_{16}$), exact $\mathcal{A}_G \cong (\mathcal{S}_2)^2$, to trivial group (i.e., no symmetry).

$$
f_{\mathcal{S}_{16}} = w\, \mathbf{I}_{16} + w'(\mathbf{1} - \mathbf{I}_{16}), \quad \text{(2 parameters)}; \tag{23}
$$

$$
f_{\mathcal{A}_G} = W_e \oplus W_2 \oplus W_3, \quad \text{(118 parameters on the isotypic components)}; \tag{24}
$$

$$
f_{\text{triv.}} = W, \quad \text{(256 parameters)}. \tag{25}
$$

To parameterize linear equivariant function $f : \mathbb{R}^{16\times d} \to \mathbb{R}^{16\times d'}$, we proceed by decoupling the input space into $\mathbb{R}^{10\times d}, \mathbb{R}^{3\times d}, \mathbb{R}^{3\times d}$ and the output space into $\mathbb{R}^{10\times d'}, \mathbb{R}^{3\times d'}, \mathbb{R}^{3\times d'}$. Now the learnable weight matrices for multidimensional input/output become $W_e \in \mathbb{R}^{10d\times 10d'}, W_2 \in \mathbb{R}^{3d\times 3d'}, W_3 \in \mathbb{R}^{3d\times 3d'}$. The construction is summarized in Algorithm 2.

Algorithm 2 Equivariant layer $f_{\mathcal{A}_G} : \mathbb{R}^{16\times d} \to \mathbb{R}^{16\times d'}$ for $\mathcal{A}_G \cong (\mathcal{S}_2)^2$

Require: The basis $B \in \mathbb{R}^{16\times 16}$ in (21) for isotypic decomposition of $\mathcal{A}_G = (\mathcal{S}_2)^2$, input $h^{(l)} \in \mathbb{R}^{16\times d}$.

Initialize: The learnable weights $W_e^{(l)} \in \mathbb{R}^{10d'\times 10d}; W_2^{(l)}, W_3^{(l)} \in \mathbb{R}^{3d'\times 3d}; M^{(l)} \in \mathbb{R}^{16\times 16}$.

1. Project $h^{(l)}$ to the isotypic component: $z^{(l)} = B^\top h^{(l)}$;
2. Perform block-wise linear transformation:
   - $z_e = W_e \ \mathrm{flatten}(z^{(l)}_{[:,:10]})$
   - $z_2 = W_2 \ \mathrm{flatten}(z^{(l)}_{[:,10:13]})$
   - $z_3 = W_3 \ \mathrm{flatten}(z^{(l)}_{[:,13:]})$
   - $z^{(l+1)} = \mathrm{concat}[z_e;\ z_2;\ z_3] \in \mathbb{R}^{16\times d'}$
3. Project back to the standard basis: $\bar{h}^{(l+1)} = B\, z^{(l+1)}$.
4. Perform pointwise nonlinearity: $h^{(l+1)} = \sigma(\bar{h}^{(l+1)})$.

return $h^{(l+1)}$

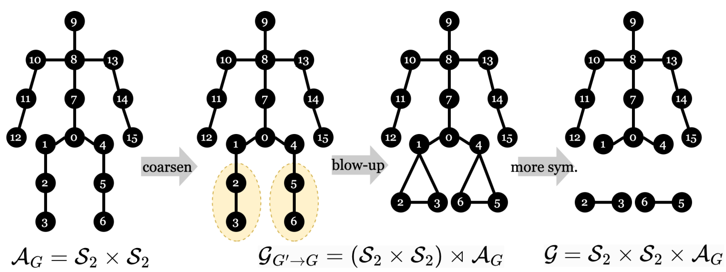


Figure 3: Human skeleton graph G, its coarsened graph G' (clustering leg joints), and blow-up of G'

C Proofs of Our Theoretical Results

C.1 Proofs of Generalization with Exact Symmetries

Lemma 1 (Risk Gap)**.** Let $\mathcal{X} = \mathbb{R}^{N\times d}, \mathcal{Y} = \mathbb{R}^{N\times k}$ be the input and output graph signal spaces on a fixed graph G. Let $X \sim \mu$ where μ is a $\mathcal{S}_N$-invariant distribution on $\mathcal{X}$. Let $Y = f^*(X) + \xi$, where $\xi \in \mathbb{R}^{N\times k}$ is random, independent of X with zero mean and finite variance and $f^* : \mathcal{X} \to \mathcal{Y}$ is $\mathcal{A}_G$-equivariant. Then, for any $f \in V$ and for any compact group $\mathcal{G} \subseteq \mathcal{S}_N$, we can decompose it as

$$f = \bar{f}_{\mathcal{G}} + f_{\mathcal{G}}^{\perp},$$

where $\bar{f}_{\mathcal{G}} = \mathcal{Q}_{\mathcal{G}} f, f_{\mathcal{G}}^{\perp} = f - \bar{f}_{\mathcal{G}}$. Moreover, the risk gap satisfies

$$\Delta(f, \bar{f}_{\mathcal{G}}) := \mathbb{E}\left[\|Y - f(X)\|_F^2\right] - \mathbb{E}\left[\|Y - \bar{f}_{\mathcal{G}}(X)\|_F^2\right] = \underbrace{-2\langle f^*, f_{\mathcal{G}}^{\perp}\rangle_\mu}_{\text{mismatch}} + \underbrace{\left\|f_{\mathcal{G}}^{\perp}\right\|_\mu^2}_{\text{constraint}}.$$

Lemma 1 is a straightforward extension of Lemma 6 in [8], which makes use of Lemma 1 in [8].

Lemma 1 in [8]. Let U be any subspace of V that is closed under $\mathcal{Q}$. Define the subspaces S and A of, respectively, the $\mathcal{G}$-symmetric and $\mathcal{G}$-anti-symmetric functions in U : $S = \{f \in U : f$ is $\mathcal{G}$-equivariant $\}$ and $A = \{f \in U : \mathcal{Q}f = 0\}$. Then U admits admits an orthogonal decomposition into symmetric and anti-symmetric parts

$$U = S \oplus A$$

Proof. The first part of Lemma 1 $f = \bar{f}_{\mathcal{G}} + f_{\mathcal{G}}^{\perp}$ follows from Lemma 1 in [8]. For the second part, by the assumption that the noise ξ is independent of X with zero mean and finite variance, we can simplify the risk gap as

$$\begin{aligned}\Delta(f, \bar{f}_{\mathcal{G}}) &:= \mathbb{E}\left[\|Y - f(X)\|_F^2\right] - \mathbb{E}\left[\|Y - \bar{f}_{\mathcal{G}}(X)\|_F^2\right] \\ &= \mathbb{E}\left[\|f^*(X) - f(X)\|_F^2\right] - \mathbb{E}\left[\|f^*(X) - \bar{f}_{\mathcal{G}}(X)\|_F^2\right]. \end{aligned} \tag{26}$$

Substituting $f = \bar{f}_{\mathcal{G}} + f_{\mathcal{G}}^{\perp}$ yields

$$\begin{aligned} &\mathbb{E}\left[\|f^*(X) - \bar{f}_{\mathcal{G}}(X) - f_{\mathcal{G}}^{\perp}(X)\|_F^2\right] - \mathbb{E}\left[\|f^*(X) - \bar{f}_{\mathcal{G}}(X)\|_F^2\right] \\ = &-2\langle f^*(X) - \bar{f}_{\mathcal{G}}(X), f_{\mathcal{G}}^{\perp}(X)\rangle_\mu + \mathbb{E}\left[\|f_{\mathcal{G}}^{\perp}(X)\|_F^2\right] \\ = &-2\langle f^*, f_{\mathcal{G}}^{\perp}\rangle_\mu + \left\|f_{\mathcal{G}}^{\perp}\right\|_\mu^2. \end{aligned} \tag{27}$$

$\square$

We remark that Lemma 6 in [8] assumes that f^* is $\mathcal{G}$-equivariant, so the first term in (27) vanishes. We are motivated from the symmetry model selection problem, and thereby relax the assumption of the chosen symmetry group $\mathcal{G}$ can differ from the target symmetry group $\mathcal{A}_{\mathcal{G}}$.

Theorem 2 (Bias-Variance-Tradeoff)**.** Let $\mathcal{X} = \mathbb{R}^{N\times d}, \mathcal{Y} = \mathbb{R}^{N\times k}$ be the graph signals spaces on a fixed graph G. Let $\mathcal{G} \subseteq \mathcal{S}_N$ with orthogonal representations ϕ on $\mathcal{X}$ and ψ on $\mathcal{Y}$. Let $X_{[i,j]} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_X^2\right)$ and $Y = f^*(X) + \xi$ where $f^*(x) = \Theta^\top x$ is $\mathcal{A}_G$-equivariant and $\Theta \in \mathbb{R}^{Nd\times Nk}$. Assume $\xi_{[i,j]}$ is random, independent of X, with mean 0 and $\mathbb{E}\left[\xi\xi^\top\right] = \sigma_\xi^2 < \infty$. Let $\hat{\Theta}$ be the least-squares estimate of Θ from n i.i.d. examples $\{(X_i, Y_i) : i = 1, \ldots, n\}$, $\Psi_{\mathcal{G}}(\hat{\Theta})$ be its equivariant version with respect to $\mathcal{G}$. Let $(\chi_\psi \mid \chi_\phi) = \int_{\mathcal{G}} \chi_\psi(g)\chi_\phi(g)\mathrm{d}\lambda(g)$ denote the scalar product of the characters. If $n > Nd + 1$ the risk gap is

$$\mathbb{E}\left[\Delta\left(f_{\hat{\Theta}}, f_{\Psi_{\mathcal{G}}(\hat{\Theta})}\right)\right] = \underbrace{-\sigma_X^2 \left\|\Psi_{\mathcal{G}}^{\perp}(\Theta)\right\|_F^2}_{\text{bias}} + \underbrace{\sigma_\xi^2 \frac{N^2dk - (\chi_\psi \mid \chi_\phi)}{n - Nd - 1}}_{\text{variance}}.$$

Theorem 2 presents the risk gap in expectation, which follows from Lemma 1, by taking f as the least-squares estimator and using assumptions in the linear regression setting. To this end, we denote $\boldsymbol{X} \in \mathbb{R}^{n\times Nd}, \boldsymbol{Y} \in \mathbb{R}^{n\times Nk}, \boldsymbol{\xi} \in \mathbb{R}^{n\times Nk}$ as the training data arranged in matrix form, where $\boldsymbol{Y} = f^*(\boldsymbol{X}) + \boldsymbol{\xi}$. Recall that the least-squares estimator of Θ in the classic regime ($n > d$) is given by

$$\hat{\Theta} := (\boldsymbol{X}^\top\boldsymbol{X})^\dagger\boldsymbol{X}^\top\boldsymbol{Y} \overset{a.e.}{=} \Theta + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\xi}, \tag{28}$$

while its equivariant map is

$$\Psi_{\mathcal{G}}(\hat{\Theta}) = \int_{\mathcal{G}} \phi(g)\, \hat{\Theta}\, \psi\left(g^{-1}\right) \mathrm{d}\lambda(g). \tag{29}$$

Our proof makes use of the following results in [8], which we restate adapted versions here for our setting.

Proposition 11 in [8]. Let $V = \{f_W : f_W(x) = W^\top x, W \in \mathbb{R}^{d\times k}, x \in \mathbb{R}^d\}$ denote the space of linear functions. Let $X \sim \mu$ with $\mathbb{E}[XX^\top] = \Sigma$. For any linear functions $f_{W_1}, f_{W_2} \in V$, the inner product on V satisfies

$$\langle f_{W_1}, f_{W_2}\rangle_\mu = \mathrm{Tr}(W_1^\top \Sigma W_2). \tag{30}$$

Theorem 13 in [8] (Simplified, Adapted)**.** Consider the same setting as Theorem 2. For $n > Nd + 1$,

$$\sigma_X^2 \mathbb{E}\left[\left\|\Psi_{\mathcal{G}}^{\perp}\left(\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{+}\boldsymbol{X}^\top\boldsymbol{\xi}\right)\right\|_F^2\right] = \sigma_\xi^2 \frac{N^2dk - (\chi_\psi \mid \chi_\phi)}{n - Nd - 1}.$$

Proof. We first plug in the least-squares expressions $\hat{\Theta}, \Psi_{\mathcal{G}}(\hat{\Theta})$ to Lemma 1 and treat the mismatch term and constraint term separately; We complete the proof by collecting common terms together.

For the mismatch term, our goal is to compute

$$-2\,\mathbb{E}\,[\langle \Theta, \hat{\Theta} - \Psi_{\mathcal{G}}(\hat{\Theta})\rangle_{\mu}], \tag{31}$$

where the expectation is taken over the test point X and the training data $\boldsymbol{X}, \boldsymbol{\xi}$.

To that end, we write

$$\left(\hat{\Theta} - \Psi_{\mathcal{G}}(\hat{\Theta})\right) x \overset{a.e.}{=} \Theta^{\top} x + \boldsymbol{\xi}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}x - \int_{\mathcal{G}} \psi(g^{-1})\left(\Theta^{\top} + \boldsymbol{\xi}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\right)\phi(g)x\,\mathrm{d}\lambda(g). \tag{32}$$

Taking expectation yields

$$\begin{aligned}\mathbb{E}_{X,\boldsymbol{X},\boldsymbol{\xi}}\,[\langle \Theta, \hat{\Theta} - \Psi_{\mathcal{G}}(\hat{\Theta})\rangle_{\mu}] &= \|\Theta\|_{\mu}^{2} + \mathbb{E}_{X,\boldsymbol{X},\boldsymbol{\xi}}\left[\langle \Theta^{\top}\boldsymbol{X}, \boldsymbol{\xi}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}x\rangle\right] \\ &\quad - \mathbb{E}_{X,\boldsymbol{X},\boldsymbol{\xi}}\left[\langle \Theta^{\top}x, \int_{\mathcal{G}} \psi(g^{-1})\left(\Theta^{\top} + \boldsymbol{\xi}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\right)\phi(g)x\,\mathrm{d}\lambda(g)\rangle\right].\end{aligned} \tag{33}$$

Note that $\boldsymbol{\xi}$ is independent with $\boldsymbol{X}$ and mean 0, so the second term in (33) vanishes. Similarly, the part $\mathbb{E}_{X,\boldsymbol{X},\boldsymbol{\xi}} \int_{\mathcal{G}} \psi(g^{-1})\left(\boldsymbol{\xi}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\right)\phi(g)x\,\mathrm{d}\lambda(g)$ also vanishes (by first taking conditional expectation of $\boldsymbol{\xi}$ conditioned on $\boldsymbol{X}$). Thus, we arrive at

$$\begin{aligned}\mathbb{E}\,[\langle \Theta, \hat{\Theta} - \Psi_{\mathcal{G}}(\hat{\Theta})\rangle_{\mu}] &= \|\Theta\|_{\mu}^{2} - \mathbb{E}_{x}\left[\langle \Theta^{\top}x, \int_{\mathcal{G}} \psi(g^{-1})\,\Theta^{\top}\,\phi(g)x\,\mathrm{d}\lambda(g)\rangle\right] \\ &= \|\Theta\|_{\mu}^{2} - \langle \Theta, \Psi_{\mathcal{G}}(\Theta)\rangle_{\mu} \\ &= \|\Psi_{\mathcal{G}}^{\perp}(\Theta)\|_{\mu}^{2} \\ &= -2\,\sigma_{X}^{2}\|\Psi_{\mathcal{G}}^{\perp}(\Theta)\|_{F}^{2},\end{aligned} \tag{34}$$

where the last equality follows from Proposition 11 in [8] with the assumption that $\Sigma = \sigma_X^2$. This finishes the computation for the mismatch term.

Now for the constraint term, we have

$$\begin{aligned}\|f_{\mathcal{G}}^{\perp}\|_{\mu}^{2} &= \|\Psi_{\mathcal{G}}^{\perp}(\hat{\Theta})\|_{\mu}^{2} &(35)\\ &= \sigma_{X}^{2}\,\mathbb{E}_{\boldsymbol{X},\boldsymbol{\xi}}\|\Psi_{\mathcal{G}}^{\perp}\left(\Theta + (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{\xi}\right)\|^{2} &(36)\\ &= \sigma_{X}^{2}\|\Psi_{\mathcal{G}}^{\perp}(\Theta)\|_{F}^{2} + \sigma_{X}^{2}\,\mathbb{E}_{\boldsymbol{X},\boldsymbol{\xi}}\|\Psi_{\mathcal{G}}^{\perp}\left((\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{\xi}\right)\|^{2}, &(37)\end{aligned}$$

where the last equality follows from linearity of expectation, $\mathbb{E}[\boldsymbol{\xi}] = 0$ and $\boldsymbol{\xi}$ independent of x.

Combining the mismatch term in (34) with the constraint term in (37), the risk gap becomes

$$\mathbb{E}\left[\Delta\left(f_{\hat{\Theta}}, f_{\Psi_{\mathcal{G}}(\hat{\Theta})}\right)\right] = -\sigma_{X}^{2}\|\,\Psi_{\mathcal{G}_L}^{\perp}(\Theta)\|^{2} + \sigma_{X}^{2}\,\mathbb{E}_{\boldsymbol{X},\boldsymbol{\xi}}\|\Psi_{\mathcal{G}_L}^{\perp}\left((\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{\xi}\right)\|^{2}, \tag{38}$$

Applying Theorem 13 in [8], the second term in (38) reduces to

$$\sigma_{X}^{2}\,\mathbb{E}_{\boldsymbol{X},\boldsymbol{\xi}}\|\Psi_{\mathcal{G}_L}^{\perp}\left((\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{\xi}\right)\|^{2} = \sigma_{\xi}^{2}\frac{N^{2}dk - (\chi_{\psi} \mid \chi_{\phi})}{n - Nd - 1}, \tag{39}$$

from which the theorem follows immediately.

□

Finally, we state a well-known result for the risk of (Ordinary) Least-Squares Estimator [2] (see [70, 71] and references therein).

Lemma 6 (Risk of Least-Squares Estimator)**.** Consider the same set-up as Theorem 2. For $n > Nd + 1$,

$$\mathbb{E}\left[\|Y - \hat{\Theta}^{\top}X\|_{F}^{2}\right] = \sigma_{\xi}^{2}\frac{Nd}{n - Nd - 1} + \sigma_{\xi}^{2}.$$

[2] In the main paper, the irreducible error term σ_ξ^2 is missing. We fix this in the Appendix and the revised version. The risk gain is of a factor $\frac{N^2dk-(\chi_\psi|\chi_\phi)}{n-1}$.

Proof. Recall X, Y denote the test sample. We denote the risk of the least-squares estimator *conditional on the training data* $\boldsymbol{X} \in \mathbb{R}^{n \times Nd}$ as $\mathcal{R}(\hat{\Theta} \mid \boldsymbol{X})$, which has the following bias-variance decomposition:

$$\mathcal{R}(\hat{\Theta} \mid \boldsymbol{X}) = \mathbb{E}\left[\|Y - \hat{\Theta}^\top X\|_F^2 \mid \boldsymbol{X}\right] \tag{40}$$

$$= \mathbb{E}\left[\|\Theta^\top X + \xi - \hat{\Theta}^\top X\|_F^2 \mid \boldsymbol{X}\right] \tag{41}$$

$$= \mathbb{E}\left[\|(\Theta - \hat{\Theta})^\top X\|_F^2 \mid \boldsymbol{X}\right] + \sigma_\xi^2, \tag{42}$$

where the last equality follows from ξ being zero mean and independent with X. The second term σ_ξ^2 is also known as *irreducible error*. We decompose the first term into

$$\mathbb{E}\left[\|(\Theta - \hat{\Theta})^\top X\|_F^2 \mid \boldsymbol{X}\right] = \mathbb{E}\left[\|(\Theta - \mathbb{E}[\hat{\Theta}])^\top X\|_F^2 + \|(\mathbb{E}[\hat{\Theta}] - \hat{\Theta})^\top X\|_F^2 \mid \boldsymbol{X}\right]. \tag{43}$$

Recall that $\hat{\Theta} \overset{a.e.}{=} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\Theta + \xi) = \Theta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \xi$. Thus $\mathbb{E}[\hat{\Theta}] = \Theta$ and (43) simplifies to $\mathbb{E}\left[\|(\mathbb{E}[\hat{\Theta}] - \hat{\Theta})^\top X\|_F^2 \mid \boldsymbol{X}\right]$.

We finish computing the risk by taking expectation over $\boldsymbol{X}$, and using $\mathbb{E}[\hat{\Theta}] - \hat{\Theta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \xi$,

$$\mathbb{E}\left[\|Y - \hat{\Theta}^\top X\|_F^2\right] = \mathbb{E}\left[\mathcal{R}(\hat{\Theta} \mid \boldsymbol{X})\right] \tag{44}$$

$$= \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{X,\xi}\left[\|(\mathbb{E}[\hat{\Theta}] - \hat{\Theta})^\top X\|_F^2 \mid \boldsymbol{X}\right]\right] + \sigma_\xi^2 \tag{45}$$

$$= \mathbb{E}\left[\|\left((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \xi\right)^\top X\|_F^2\right] + \sigma_\xi^2 \tag{46}$$

$$= \sigma_\xi^2 \operatorname{tr}\left(\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}]\, \sigma_X^2 I\right) + \sigma_\xi^2. \tag{47}$$

By [72, Lemma 2.3], for $n > Nd + 1$, $\mathbb{E}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}] = \frac{Nd}{n - Nd - 1} I$. Putting this in (47) completes the proof. $\square$

C.2 Proofs of Generalization with Approximate Symmetries

Corollary 3 (Risk Gap via Graph Coarsening)**.** Let $\mathcal{X} = \mathbb{R}^{N \times d}, \mathcal{Y} = \mathbb{R}^{N \times k}$ be the input and output graph signal spaces on a fixed graph G. Let $X \sim \mu$ where μ is a $\mathcal{S}_N$-invariant distribution on $\mathcal{X}$. Let $Y = f^*(X) + \xi$, where $\xi \in \mathbb{R}^{N \times k}$ is random, independent of X with zero mean and finite variance, and $f^* : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times k}$ be an approximately equivariant mapping with equivariance rate κ. Then, for any G' that coarsen G up to error ϵ, for any $f \in V$, we have

$$\Delta(f, \bar{f}_{\mathcal{G}_{G \to G'}}) = \underbrace{-2\langle f^*, f^\perp_{\mathcal{G}_{G \to G'}}\rangle_\mu}_{\text{mismatch}} + \underbrace{\left\|f^\perp_{\mathcal{G}_{G \to G'}}\right\|_\mu^2}_{\text{constraint}} \geq (1 - 2\kappa(\epsilon)) \left\|f^\perp_{\mathcal{G}_{G \to G'}}\right\|_\mu^2.$$

Proof. We start by simplifying the mismatch term in Lemma 1,

$$\begin{aligned}
-2\mathbb{E}\left[\langle f^*(x), f^\perp_{\mathcal{G}_{G \to G'}}(x)\rangle\right] &= -2\mathbb{E}\left[\langle f^*(x) - f^*_{\mathcal{G}_{G \to G'}}(x) + f^*_{\mathcal{G}_{G \to G'}}(x), f^\perp_{\mathcal{G}_{G \to G'}}(x)\rangle\right] \\
&= -2\mathbb{E}\left[\langle \underbrace{f^*(x) - f^*_{\mathcal{G}_{G \to G'}}(x)}_{\mathcal{G}_L\text{-anti-symmetric part of } f^*}, \underbrace{f^\perp_{\mathcal{G}_{G \to G'}}(x)}_{\mathcal{G}_L\text{-anti-symmetric part of } f} \rangle\right] \\
&\geq -2\,\|f^* - f^*_{\mathcal{G}_{G \to G'}}\|_\mu \,\|f^\perp_{\mathcal{G}_{G \to G'}}\|_\mu \quad \text{(By Cauchy Schwarz Ineq.)} \\
&\geq -2\,\kappa(\epsilon)\,\|f^\perp_{\mathcal{G}}\|_\mu. \quad \text{(By Definition 4 Approx. Equiv. Map)}
\end{aligned}$$

Putting this together with the constraint term completes the proof. $\square$

Corollary 4 (Bias-Variance-Tradeoff via Graph Coarsening)**.** Consider the same linear regression setting in Theorem 2, except now f^* is an approximately equivariant mapping with equivariance rate κ, and $\mathcal{G} = \mathcal{G}_{G\to G'}$ is controlled by G' that coarsens G up to error ϵ. Denote the orthogonal representations of $\mathcal{G}_{G\to G'}$ on $\mathcal{X}, \mathcal{Y}$ as ϕ', ψ', respectively. Let $(\chi_{\psi'} \mid \chi_{\phi'}) = \int_{\mathcal{G}_{G\to G'}} \chi_{\psi'}(g)\chi_{\phi'}(g)\mathrm{d}\lambda(g)$ denote the scalar product of the characters. If $n > Nd+1$ the risk gap is bounded by

$$\mathbb{E}\left[\Delta\left(f_{\hat{\Theta}}, f_{\Psi_{\mathcal{G}_{G\to G'}}(\hat{\Theta})}\right)\right] \geq (1-2\kappa(\epsilon))\,\sigma_\xi^2 \frac{N^2dk - (\chi_{\psi'} \mid \chi_{\psi'})}{n - Nd - 1}.$$

Proof. It follows immediately from applying Theorem 13 in [8] to Corollary 3 with $\mathcal{G} = \mathcal{G}_{G\to G'}$. $\square$

D Example Details

D.1 Example 3.1

Consider $\mathcal{G} = \mathcal{S}_3, \mathcal{G} = \mathcal{S}_2, \mathcal{X} = \mathbb{R}^3, \mathcal{Y} = \mathbb{R}^3$, and $x \sim \mathcal{N}(0, \sigma_X^2 I_d)$. The target function is linear, i.e., $f^*(x) = \Theta^\top x$ for some $\Theta \in \mathbb{R}^{3\times 3}$. In other words, we are learning linear functions on a fixed graph domain with 3 nodes. Suppose the target function is $\mathcal{S}_2$-equivariant such that it has the form

$$\Theta = \begin{bmatrix} a & b & c \\ b & a & c \\ d & d & e \end{bmatrix}, \quad a, b, c, d, e \in \mathbb{R}. \tag{48}$$

Now, we project Θ in (48) to $\mathcal{S}_3$-equivariant space using the intertwine average 5 with the orthogonal representation of $\mathcal{S}_3$. Direct calculation yields

$$\Psi_{\mathcal{S}_3}(\Theta) = \begin{bmatrix} \frac{1}{3}(2a+e) & \frac{1}{3}(b+c+d) & \frac{1}{3}(b+c+d) \\ \frac{1}{3}(b+c+d) & \frac{1}{3}(2a+e) & \frac{1}{3}(b+c+d) \\ \frac{1}{3}(b+c+d) & \frac{1}{3}(b+c+d) & \frac{1}{3}(2a+e) \end{bmatrix} \tag{49}$$

$$\Psi_{\mathcal{S}_3}^{\perp}(\Theta) = \Theta - \Psi_{\mathcal{S}_3}(\Theta) = \begin{bmatrix} \frac{1}{3}(a-e) & \frac{1}{3}(2b-c-d) & \frac{1}{3}(-b+2c-d) \\ \frac{1}{3}(2b-c-d) & \frac{1}{3}(a-e) & \frac{1}{3}(-b+2c-d) \\ \frac{1}{3}(-b-c+2d) & \frac{1}{3}(-b-c+2d) & \frac{1}{3}(-2a+2e). \end{bmatrix} \tag{50}$$

Therefore, the bias term evaluates to

$$-\sigma_X^2 \,\|\Psi_{\mathcal{S}_3}^{\perp}(\Theta)\|^2 = -\sigma_X^2\left(\frac{2(a-e)^2}{3} + \frac{2(-2b+c+d)^2}{9} + \frac{2(b-2c+d)^2}{9} + \frac{2(b+c-2d)^2}{9}\right). \tag{51}$$

For the variance term, recall $\chi_{\psi_{\mathcal{S}_3}}, \psi_{\mathcal{S}_3}$ are both the standard representation of $\mathcal{S}_3$, we have

$$\left(\chi_{\psi_{\mathcal{S}_3}} \mid \chi_{\phi_{\mathcal{S}_3}}\right) = \frac{1}{6}(3^2+1^2+1^2+1^2+0^2+0^2) = 2. \tag{52}$$

Therefore, the variance term evaluates to

$$\sigma_\xi^2 \frac{N^2 - (\chi_\psi \mid \chi_\psi)}{n-N-1} = \sigma_\xi^2 \frac{7}{n-4}. \tag{53}$$

Putting (51) and (53) together yields the generalization gap of for the least square estimator $f_{\hat{\Theta}}$ compared to its $\mathcal{S}_3$-equivariant version $f_{\Psi_{\mathcal{S}_3}(\hat{\Theta})}$.

As a comparison, when choosing the symmetry group of the target function $\mathcal{G} = \mathcal{S}_2$, the bias vanishes and note that $\left(\chi_{\psi_{\mathcal{S}_2}} \mid \chi_{\phi_{\mathcal{S}_2}}\right) = \frac{1}{2}(3^2+1^2) = 5$, so generalization gap is

$$\mathbb{E}\left[\Delta\left(f_{\hat{\Theta}}, f_{\Psi_{\mathcal{S}_2}(\hat{\Theta})}\right)\right] = \sigma_\xi^2 \frac{4}{n-4}. \tag{54}$$

We see that choosing $\mathcal{G} = \mathcal{S}_3$ is better if $a \approx e, b \approx c \approx d$ (i.e., f^* is approximately $\mathcal{S}_3$-invariant) and the training sample size n small, whereas $\mathcal{S}_2$ is better vice versa. This analysis illustrates the advantage of choosing a (suitably) larger symmetry group to induce a smaller hypothesis class when learning with limited data, and introduce useful inductive bias when the target function is approximately symmetric with respect to a larger group. We further validate our theoretical analysis via simulation, with details and results shown in Figure 4.

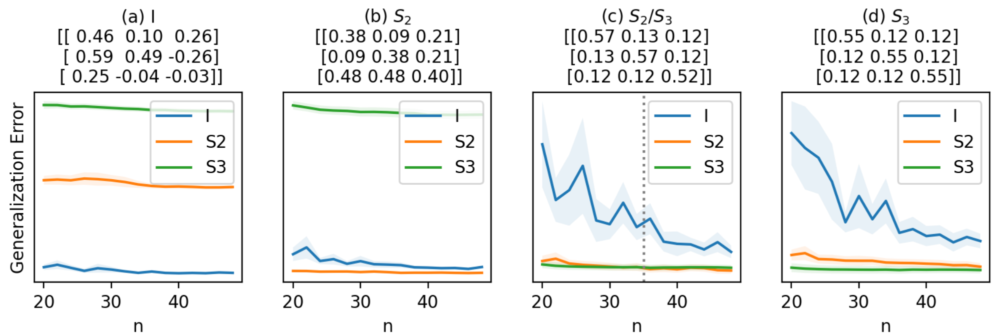


Figure 4: Choosing the symmetry group corresponding to the target function usually yields the best generalization ((a), (b), (d)), but not always: when the number of training data n is small and the target function f is approximately equivariant with respect to a larger group, choosing the larger symmetry group could yield further generalization gain, as shown in (c) empirically. Dashed gray vertical line highlights the theoretical threshold $n^* \approx 35$, before which using $\mathcal{S}_3$ yields better generalization than $\mathcal{S}_2$, validating our theoretical analysis. We set $\sigma_X^2 = 1, \sigma_\xi^2 = \frac{1}{64}$, conduct 10 random runs and compute the generalization error based on 300 test points. We obtain the estimators via stochastic gradient descent, and enforce symmetry via tying weights. Titles of each subplot indicate the symmetry of the target function, and display the target function values.

D.2 Example: Approximately Equivariant Mapping on a Geometric Graph

In this section, we illustrate a construction of an approximately equivariant mapping. We focus on a version of Definition 3 that does not take to account the symmetries of G'. Namely, we consider a definition of the approximate symmetries as

$$\mathcal{G}_{G \to G'} := \mathcal{S}_{c_1} \times \mathcal{S}_{c_2} \ldots \times \mathcal{S}_{c_M} \subset \mathcal{S}_N.$$

Equivalently, we restrict the analysis to coarsening graphs G' that are asymmetric.

Background from graphon-signal analysis. To support our construction, we cite some definitions and results from [73].

Definition 8. *Let $r > 0$. The graphon-signal space with signals bounded by r is $\mathcal{WL}_r := \mathcal{W} \times L_r^\infty[0,1]$, where $L_r^\infty[0,1]$ is the ball of radius r in $L^\infty[0,1]$. The distance in $\mathcal{WL}_r$ is defined for $(W,s),(V,g) \in \mathcal{WL}_r$ by*

$$d_\square\big((W,s),(V,g)\big) := \|(W,s)-(V,g)\|_\square := \|W-V\|_\square + \|s-g\|_1.$$

Moreover,

$$\delta_\square\big((W,s),(V,g)\big) = \inf_\phi d_\square\big((W,s),(V^\phi,g^\phi)\big),$$

where $g^\phi(x) = g(\phi(x))$ and ϕ is a measure preserving bijection.

Any graph-signal induces a graphon signal in the natural way, as in Definition 1. The cut norm and distance between to graph-signals is defined to be the cut norm and distance between the two induced graphon-siganl respectively. Similarly, the L_1 distance between a signal q on a graph and a signal s on $[0,1]$ is defined to be the L_1 distance between the induced signal from q and s.

The supremum in the definition of cut distance between two induced graphon-signals is realized for some measure preserving bijection.

Sampling graphon-signals. The following construction is from [73, Section 3.4]. Let $\Lambda = (\lambda_1, \ldots \lambda_N) \in [0,1]^N$ be N independent uniform random samples from $[0,1]$, and $(W,s) \in \mathcal{WL}_r$. We define the *random weighted graph* $W(\Lambda)$ as the weighted graph with N nodes and edge weight $w_{i,j} = W(\lambda_i, \lambda_j)$ between node i and node j. We similarly define the *random sampled signal* $s(\Lambda)$ with value $s_i = s(\lambda_i)$ at each node i. Note that $W(\Lambda)$ and $s(\Lambda)$ share the sample points Λ. We then define a random simple graph as follows. We treat each $w_{i,j} = W(\lambda_i, \lambda_j)$ as the parameter of a

Bernoulli variable $e_{i,j}$, where $\mathbb{P}(e_{i,j} = 1) = w_{i,j}$ and $\mathbb{P}(e_{i,j} = 0) = 1 - w_{i,j}$. We define the *random simple graph* $\mathbb{G}(W, \Lambda)$ as the simple graph with an edge between each node i and node j if and only if $e_{i,j} = 1$.

The following theorem is from [73, Theorem 3.6]

Theorem 1 (Sampling lemma for graphon-signals)**.** *Let $r > 1$. There exists a constant $N_0 > 0$ that depends on r, such that for every $N \geq N_0$, every $(W, s) \in \mathcal{WL}_r$, and for $\Lambda = (\lambda_1, \ldots \lambda_N) \in [0,1]^N$ independent uniform random samples from $[0,1]$, we have*

$$\mathbb{E}\Big(\delta_{\square}\Big(\big(W, s\big), \big(\mathbb{G}(W, \Lambda), s(\Lambda)\big)\Big)\Big) < \frac{15}{\sqrt{\log(N)}}. \tag{55}$$

By Markov's inequality and (55), for any $0 < p < 1$, there is an event of probability $1 - p$ (regarding the choice of Λ) in which

$$\delta_{\square}\Big(\big(W, s\big), \big(\mathbb{G}(W, \Lambda), s(\Lambda)\big)\Big) < \frac{15}{p\sqrt{\log(N)}}. \tag{56}$$

Stability to deformations of mappings on geometric graphs. Let $\mathcal{M}$ be a metric space with an atomless standard probability measure defined over the Borel sets (up to completion of the measure). Such a probability space is equivalent to the standard probabiltiy space $[0,1]$ with Lebesgue measure. Namely, there are co-null sets $A \subset \mathcal{M}$ and $B \subset [0,1]$, and a measure preserving bijection $\phi : A \to B$. Hence, graphon analysis applied as-is when replacing the domain $[0,1]$ with $\mathcal{M}$.

Suppose that we are interested in a target function $f_{\mathcal{M}} : L^1(\mathcal{M}) \to L^1(\mathcal{M})$ that is stable to deformations in the following sense.

Definition 9. *Let $\epsilon > 0$. A measurable bijection $\nu : \mathcal{M} \to \mathcal{M}$ is called a* deformation up to ϵ*, if there exists an event $B_\epsilon \subset \mathcal{M}$ with probability greater than $1 - \epsilon$ such that for every $x \in B_\epsilon$*

$$d_{\mathcal{M}}\big(\nu(x), x\big) < \epsilon.$$

The mapping $f_{\mathcal{M}} : L^2(\mathcal{M}) \to L^2(\mathcal{M})$ is called stable to deformations *with* stability constant C*, if for any deformation ν up to ϵ, and every $s \in L^1(\mathcal{M})$, we have*

$$\|f_{\mathcal{M}}(s) - f_{\mathcal{M}}(s \circ \nu) \circ \nu^{-1}\|_1 < C\epsilon.$$

Suppose that we observe a discretized version of the domain $\mathcal{M}$, defined as follows. There is a graphon $W : \mathcal{M}^2 \to [0,1]$ defined as

$$W(x, y) = r\big(d(x, y)\big), \tag{57}$$

where $r : \mathbb{R}_+ \to [0,1]$ is a decreasing function with support $[0, \rho]$. Instead of observing W, we observe a weighted graph $G = \mathbb{G}(W, \Lambda)$ with node set $[N]$, sampled from W on the random independent points $\Lambda = \{\lambda_n\}_{n=1}^N \subset \mathcal{M}$ as above. Suppose moreover that any graph signal is sampled from a signal in $L^1(\mathcal{M})$, on the same random points Λ, as above.

Suppose that the target $f_{\mathcal{M}}$ on the continuous domain is well approximated by some mapping $f^* : L^2[N] \to L^2[N]$ on the discrete domain in the following sense. For every $s \in L^1(\mathcal{M})$, let s_G be the graph signal sampled on the random samples $\{\lambda_n\}_n$. Then there is an event of high probability such that

$$\|f^* s_G - \{\big(f_{\mathcal{M}}(s)\big)(x_n)\}_n\|_1 < e$$

for some small e. We hence consider f^* as the target mapping of the learning problem. One example of such a scenario is when there exists some Lipschitz continuous mapping $\Theta : \mathcal{WL}_r \to \mathcal{WL}_r$ with Lipschitz constant L, such that $f_{\mathcal{M}} = \Theta(W, \cdot)$ and $f^* = \Theta(G, \cdot)$. Indeed, by (56), for some p as small as we like, there is an event of probability $1 - p$ in which, up to a measure preserving bijection,

$$\begin{aligned}\|f_{\mathcal{M}} s - f^* s_G\|_1 &\leq \delta_{\square}\Big(\big(W, f_{\mathcal{M}} s\big), \big(G, f^* s_G\big)\Big)\\ &\leq L\delta_{\square}\Big(\big(W, s\big), \big(G, s_G\big)\Big) < \frac{15L}{p\sqrt{\log(N)}} = e.\end{aligned}$$

A concrete example is when Θ is a message passing neural network (MPNN) with Lipschitz continuous message and update functions, and normalized sum aggregation [73, Theorem 4.1].

Let G' be a graph that coarsens G up to error ϵ. In the same event as above, by (56), up to a measure preserving bijection,

$$d_{\square}(W_{G'}, W) \leq d_{\square}(W_{G'}, W_G) + d_{\square}(W_G, W) \leq \epsilon + e = u. \tag{58}$$

We next show an approximation property that we state here informally: since $W(x, y) \approx 0$ for x away from y, we must have $W_{G'}(x, y) \approx 0$ as well for a set of high measure. Otherwise, $\delta_{\square}(W_{G'}, W)$ cannot be small. By this, any approximate symmetry of G is a small deformation, and, hence, f^* is an approximately equivariant mapping.

Equivariant mappings on geometric graphs. In the following, we construct a scenario in which f^* can be shown to be approximately equivariant in a restricted sense. For simplicity, we restrict to the case $r = \mathbb{1}_{[0,\rho]}$ in the geometric graphon W of (57). Denote the induced graphon $W_{G'} = T$. Given $h > 0$, define the *h-diagonal*

$$d_h = \{(x, y) \in \mathcal{M}^2 \mid d_{\mathcal{M}}(x, y) \leq h\}.$$

In the following, all distances are assumed to be up to the best measure preserving bijection.

If there is a domain $S' \times T' \in \mathcal{M}^2$ outside the ρ-diagonal in which $T(x, y) > c$ for some $c > 0$, we must have

$$\|W - T\|_{\square} \geq \int_{S'} \int_{T'} T(x, y) dy dx = c\mu(S')\mu(T').$$

Hence, since by (58), $\|W - T\|_{\square} < u$, for every $S' \times T'$ that does not intersect d_ρ, we must have

$$\int_{S'} \int_{T'} T(x, y) dy dx \leq u.$$

In other words, for any two sets S, T with distance more than ρ ($\inf_{s \in S, t \in T} d_\mu(s, t) > \rho$), we have

$$\int_{S} \int_{T} T(x, y) dy dx \leq u.$$

This formalizes the statement "$W(x, y) \approx 0$ for x away from y" from above.

Next, we develop the analysis for the special case $\mathcal{M} = [0, 1]$ with the standard metric and Lebesgue probability measure. We note that the analysis can be extended to $\mathcal{M} = [0, 1]^D$ for a general dimension $D \in \mathbb{N}$.

For every $z \in [0, 1]$, we have

$$\int_{[z+\rho/\sqrt{2},1]} \int_{[0,z-\rho/\sqrt{2}]} T(x, y) dy dx \leq u,$$

and

$$\int_{[0,z-\rho/\sqrt{2}]} \int_{[z+\rho/\sqrt{2},1]} T(x, y) dy dx \leq u.$$

Let $\nu > 0$. We take a grid $\{x_j\} \in [0, 1]$ of spacing $\sqrt{2}\nu$. The sets

$$\bigcup_j [x_j + \rho/\sqrt{2}, 1] \times [0, x_j - \rho/\sqrt{2}] , \quad \bigcup_j [0, x_j - \rho/\sqrt{2}] \times [x_j + \rho/\sqrt{2}, 1]$$

cover d_ν^c (where d_ν^c is the complement of d_ν). Hence,

$$\iint_{d_\nu^c} T(x, y) dy dx \leq \sum_{j=1}^{1/\sqrt{2}\nu} \int_{[x_j+\rho/\sqrt{2},1]} \int_{[0,x_j-\rho/\sqrt{2}]} T(x, y) dy dx$$
$$+ \sum_{j=1}^{1/\sqrt{2}\nu} \int_{[0,x_j-\rho/\sqrt{2}]} \int_{[x_j+\rho/\sqrt{2},1]} T(x, y) dy dx$$

$$\leq \frac{2}{\sqrt{2}\nu}u.$$

We take $\frac{2}{\sqrt{2}\nu}u = t$, for $u \ll t \ll 1$, namely, $\nu = \sqrt{\frac{2u}{t}}$. For example, we may take $t = u^{1/3}$, and $\nu = \sqrt{2}u^{1/2-1/6} = \sqrt{2}u^{1/2}$, assuming that $\rho < u^{1/3}$. Hence, we have

$$\iint_{d^c_{u^{1/3}}} T(x, y) \leq \sqrt{2}u^{1/3}.$$

To conclude, the probability of having an edge between nodes λ_i and λ_j in $\overline{G'}_N$ which are further away than $u^{1/3}$, namely, $d_{\mathcal{M}}(\lambda_i, \lambda_j) > u^{1/3}$, is less than $\sqrt{2}u^{1/3}$.

Suppose that G' is asymmetric. This means that symmetries of $\mathcal{G}_{G\to G'}$ can only permute between nodes that have an edge between them in the blown-up graph $\overline{G'}_N$. The probability of having an edge between nodes further away than $u^{1/3}$ is less than $\sqrt{2}u^{1/3}$. Hence, a symmetry in $\mathcal{G}_{G\to G'}$ can be seen as a small deformation, where for each node λ_i and a random uniform $g \in \mathcal{G}_{G\to G'}$, the probability that λ_i it is mapped by g to a node of distance less than $u^{1/3}$ is more than $1 - \sqrt{2}u^{1/3}$. Any symmetry g in $\mathcal{G}_{G\to G'}$ induces a measure preserving bijection ν in $\mathcal{M} = [0, 1]$, by permuting the intervals of the partition $\mathcal{P}_N$ of Definition 1. As a result, the set of points that are mapped further away than $u^{1/3}$ under ν has probability upper bounded by $\sqrt{2}u^{1/3}$, and symmetries in $\mathcal{G}_{G\to G'}$ can be seen as a small deformation ν according to Definition 9 (in high probability). This means that

$$\|f_{\mathcal{M}}(s) - f_{\mathcal{M}}(s \circ \nu) \circ \nu^{-1}\|_1 < C\sqrt{2}u^{1/3},$$

so by the triangle inequality, we have

$$\|f^*(s_G) - g^{-1}f^*(gs_G)\|_1 < 2e + C\sqrt{2}u^{1/3} = \epsilon', \tag{59}$$

Next, we show that f^* is approximately equivariant in a restricted sense, where we limit ourselves to a symmetry group

$$\mathcal{G}_{G\to G'} = \mathcal{S}_{c_1} \times \mathcal{S}_{c_2} \ldots \times \mathcal{S}_{c_M}$$

in Definition 3, without the symmetries of $\overline{\mathcal{A}}_{G'}$.

Equation (59) leads to

$$\|f^*(s_G) - \mathcal{Q}_{\mathcal{G}_{G\to G'}}(f^*)(s_G)\|_1 = \|f^*(s_G) - \frac{1}{|\mathcal{G}_{G\to G'}|}\sum_{g\in\mathcal{G}_{G\to G'}} g^{-1}f^*(gs_G)\|_1 \tag{60}$$

$$\leq \frac{1}{|\mathcal{G}_{G\to G'}|}\sum_{g\in\mathcal{G}_{G\to G'}} \|f^*(s_G) - g^{-1}f^*(gs_G)\|_1 < \epsilon'. \tag{61}$$

Since for any $q \in L^2[0, 1] \cap L^\infty[0, 1]$ we have $\|q\|_2^2 \leq \|q\|_\infty \|q\|_1$, we can bound

$$\|f^*(s_G) - \mathcal{Q}_{\mathcal{G}_{G\to G'}}(f^*)(s_G)\|_2 < \sqrt{2\|f^*(s_G)\|_\infty}\sqrt{\epsilon'}.$$

Denote $\|f^*\|_\infty := \int \|f^*(s_G)\|_\infty d\mu(s_G)$, and suppose that $\|f^*\|_\infty$ is finite. Hence, if μ is a probability measure, we have

$$\left\|f^* - \mathcal{Q}_{\mathcal{G}_{G\to G'}}(f^*)\right\|_\mu < \sqrt{2\|f^*\|_\infty}\sqrt{\epsilon'}.$$

This shows a modified version of approximate equivariance, where the approximation rate is also a function of the size of the graph N, and goes to zero as $N \to \infty$ and $\epsilon \to 0$.

In future work, we will extend this example to more general metric space $\mathcal{M}$ and to non-trivial symmetry groups $\overline{\mathcal{A}}_{G'}$. Intuitively, most random geometric graphs are “close to asymmetric.” This means that for “most” G', the symmetries of $\overline{\mathcal{A}}_{G'}$ can only permute between nodes connected by an edge, and so are the symmetries of $\mathcal{G}_{G\to G'}$. For this, we need to extend Definition 9 by treating G' probabilistically.

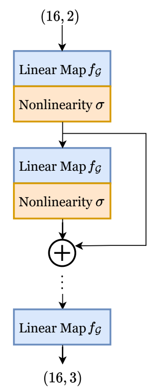
E Experiment Details

The source code will be made available in the final version of the paper. All experiments were conducted on a server with 256 GB RAM and 4 NVIDIA RTX A5000 GPU cards.

E.1 Application: Human Pose Estimation

Data. We use the standard benchmark dataset, Human3.6M [65], with the same protocol as in [66]: We train the models on 1.56M poses (from human subjects $S1, S5, S6, S7, S8$) and evaluate them on 0.54M poses (from human subjects $S9, S11$). We use the method described in [74] to normalize the inputs (2D joint poses) to $[-1, 1]$ and align the targets (3d joint poses) with the root joint.

Graph Networks with Equivariant Modules. We give detailed description of $\mathcal{G}$-Net and its variants used in the experiments. Figure inset illustrates the architecture of $\mathcal{G}$-Net. For the human skeleton graph with $N = 16$, we have $f_\mathcal{G} : \mathbb{R}^{16\times d} \to \mathbb{R}^{16\times k}$, where d, k represent the input dimension and output dimension (for each layer). Let $f_\mathcal{G}[i, j] : \mathbb{R}^{16} \to \mathbb{R}^{16}$ denote its (i, j)-th slice.



1. $\mathcal{G}$-Net with strict equivariance using equivariant linear map $f_\mathcal{G}$ (see Table 5):

- $\mathcal{S}_{16}$: $f_{\mathcal{S}_{16}}[i, j] \in \mathbb{R}^{16\times 16}$ is a diagonal matrix, with one learnable scalar a on diagonal and another learnable scalar b off diagonal.
- Relax-$\mathcal{S}_{16}$: We relax $f_{\mathcal{S}_{16}}[i, j]$ by having 16 different pairs of scalars $(a_i, b_i), i \in [16]$, such that each node i can map to itself and communicate to its neighbors in a different way (controlled by (a_i, b_i)), while still treat all neighbors equally (by using the same b_i for nodes $j \neq i$).
- $(\mathcal{S}_2)^6$: We use Algorithm 2 while replacing $\mathcal{A}_G$ with the symmetry group on a disconnected graph G_0 consists of the orbits in G, i.e. G_0 has the same nodes as G, but only retaining the edges among $(1, 4), (2, 5), (3, 6), (10, 13), (11, 14), (12, 15)$.
- $\mathcal{A}_G$: We use Algorithm 2.
- $\mathcal{S}_2$: We use Algorithm 2 while replacing $\mathcal{A}_G$ with $\mathcal{S}_2$ representing the bilateral symmetry on the human skeleton graph (i.e., the left arms and legs must flip together, similarly for the right arms and legs).
- Trivial: We allow $f[i, j] \in \mathbb{R}^{16\times 16}$ to be arbitrary, i.e., it has 16×16 learnable scalars.

We remark that for $\mathcal{S}_{16}$ and Relax-$\mathcal{S}_{16}$, we implement them by tying weights; for $(\mathcal{S}_2)^6, \mathcal{A}_G, \mathcal{S}_2$, we implement them by projecting to isotypic component as shown in Algorithm 2.

2. $\mathcal{G}$-Net augmented with graph convolution $Af_\mathcal{G}(x)$, denoted as $\mathcal{G}$-Net(gc) (see Table 5): We apply the equivariant linear map $f_\mathcal{G}$ in 1. and obtain the output $f_\mathcal{G}(x) \in \mathbb{R}^{16\times k}$; We then apply graph convolution by multiplication from the left, i.e., $Af_\mathcal{G}(x) \in \mathbb{R}^{16\times k}$.

3. $\mathcal{G}$-Net augmented with graph convolution and learnable edge weights, denoted as $\mathcal{G}$-Net(gc+ew) (see Table 1[3]): We further learn the edge weights for the adjacency matrix A, by $\texttt{softmax}(M \odot A)$ where $M \in \mathbb{R}^{16}$ represents the learnable edge weights, and $M_{i,j}$ is nonzero when $A_{i,j} \neq 0$ and 0 elsewhere. This is inspired from SemGCN [66]. Besides the groups discussed in 1., we also implemented Relax-$(\mathcal{S}_6)^2$ which corresponds to tying weights among the coarsened graph orbits, consists of 4 spline nodes (singleton orbits) and 2 orbits for the left/right arm and leg nodes.

4. $\mathcal{G}$-Net augmented with graph locality constraints $(A \odot f_\mathcal{G})(x)$, denoted as $\mathcal{G}$-Net(pt) (see Table 5): We perform pointwise multiplication $A \odot f_\mathcal{G}[i, j]$ at each (i, j)-th slice of $f_\mathcal{G}$. In practice, we also allow learnable edge weights as done in 3.

Experimental Set-up. We design $\mathcal{G}$-Net to have 4 layers (with batch normalization and residual connections in between the hidden layers), 128 hidden units, and use ReLU nonlinearity. This allows $\mathcal{G}$-Net(gc+ew) to recover SemGCN [66] when choosing $\mathcal{G} = \mathcal{S}_{16}$. We train our models for maximally 30 epochs with early stopping. For comparison purpose, we use the same optimization routines as in SemGCN [66] and perform the hyper-parameter search of learning rates $\{0.001, 0.002\}$.

Evaluation. Table 5 shows results of $\mathcal{G}$-Net and its variants when varying the choice of $\mathcal{G}$. We observe that using the automorphism group $\mathcal{A}_G$ does not give the best performance, while imposing no symmetries (Trivial) or a relaxed version of $\mathcal{S}_{16}$ yields better results.

[3] There is a typo in Table 1, where $(\mathcal{S}_2)^6$ should be corrected to Relax-$\mathcal{S}_{16}$, and $(\mathcal{S}_6)^2$ should be corrected to Relax-$(\mathcal{S}_6)^2$.

Table 5: 3D human pose prediction using $\mathcal{G}$-Net and its variants. Error ($\pm$ std) measured by Mean Per-Joint Position Error (MPJPE) and MPJPE after rigid alignment (P-MPJPE) across 3 runs. All methods use the same hidden dimension $d = 128$. Bold type indicates the top-2 performance among each variant. "NA" indicates the loss fails to converge.

MPJPE $\downarrow$	$\mathcal{S}_{16}$	Relax-$\mathcal{S}_{16}$	$(\mathcal{S}_2)^6$	$\mathcal{A}_G = (\mathcal{S}_2)^2$	$\mathcal{S}_2$	Trivial
$\mathcal{G}$-Net	NA	$\mathbf{47.97 \pm 0.47}$	52.97 ± 0.79	48.30 ± 0.69	48.95 ± 0.31	$\mathbf{42.86 \pm 0.64}$
$\mathcal{G}$-Net(gc)	NA	54.50 ± 4.33	52.97 ± 0.64	49.40 ± 1.37	$\mathbf{48.72 \pm 0.39}$	$\mathbf{43.24 \pm 0.82}$
$\mathcal{G}$-Net(pt)	41.54 ± 0.47	$\mathbf{40.44 \pm 0.61}$	52.47 ± 0.48	40.63 ± 0.26	48.19 ± 0.13	$\mathbf{38.41 \pm 0.31}$
P-MPJPE $\downarrow$	$\mathcal{S}_{16}$	Relax-$\mathcal{S}_{16}$	$(\mathcal{S}_2)^6$	$\mathcal{A}_G = (\mathcal{S}_2)^2$	$\mathcal{S}_2$	Trivial
$\mathcal{G}$-Net	NA	$\mathbf{36.45 \pm 0.56}$	41.66 ± 0.28	37.17 ± 0.59	37.27 ± 0.27	$\mathbf{32.59 \pm 0.62}$
$\mathcal{G}$-Net(gc)	NA	40.61 ± 0.99	41.87 ± 0.80	37.62 ± 1.32	$\mathbf{36.97 \pm 0.78}$	$\mathbf{33.05 \pm 0.81}$
$\mathcal{G}$-Net(pt)	32.31 ± 0.03	$\mathbf{31.11 \pm 0.68}$	41.45 ± 0.28	31.35 ± 0.14	37.56 ± 0.12	$\mathbf{29.68 \pm 0.22}$

Table 6: 3D human pose prediction using $\mathcal{G}$-Net(gc+ew), where the models induced from each choice of $\mathcal{G}$ are set to have roughly the same number of parameters. d denotes the number of hidden units.

$\mathcal{G}$-Net	Number of Parameters	Number of Epochs	MPJPE	P-MPJPE
$\mathcal{S}_{16}$	0.27M ($d = 128$)	50	43.48	34.96
Relax-$\mathcal{S}_{16}$	0.27M ($d = 32$)	20	**40.08**	**32.08**
$\mathcal{A}_G = (\mathcal{S}_2)^2$	0.22M ($d = 16$)	30	44.10	34.12
Trivial	0.22M ($d = 10$)	30	45.05	34.79

Additional Evaluation. Table 6 shows the experiments when we keep the number of parameters roughly the same across different choices of $\mathcal{G}$.

E.2 Application: Traffic Flow Prediction

Data. The METR-LA traffic dataset, [67], contains traffic information collected from 207 sensors in the highway of Los Angeles County from Mar 1st 2012 to Jun 30th 2012 [75]. We use the same traffic data normalization and $70/10/20$ train/validation/test data split as [67]. We consider two different traffic graphs constructed from the pairwise road network distance matrix: (1) the sensor graph G introduced in [67] based on applying a thresholded Gaussian kernel (degree distribution in Figure 5e); (2) the sparser graph G_s based on applying the binary mask where the (i, j) entry is nonzero if and only if nodes i, j lie on the same highway (degree distribution in Figure 5d). We construct the second variant to more faithfully model the geometry of the highway sensors, illustrated in Figure 5a.

Graph coarsening. We choose 2 clusters based on highway intersection and flow direction, indicated by colors (Figure 5c (b)), and 9 clusters based on highway labels (Figure 5c (c)).

Model. We use a standard baseline, DCRNN proposed in [67]. DCRNN is built on a core recurrent module, DCGRU cell, which iterates as follows: Let $x_{i,t}, h_{i,t}$ denote the i-th node feature and hidden state vector at time t; Let X_t, R_t, H_{t-1} be the matrices of stacking feature vectors $x_{i,t}, r_{i,t}, h_{i,t-1}$ as rows.

$$z_{i,t} = \sigma_g \left(W_z \, x_{i,t} + U_z \, h_{i,t-1} + b_z \right) \tag{62}$$

$$r_{i,t} = \sigma_g \left(W_r \, x_t + U_r \, h_{t-1} + b_r \right) \tag{63}$$

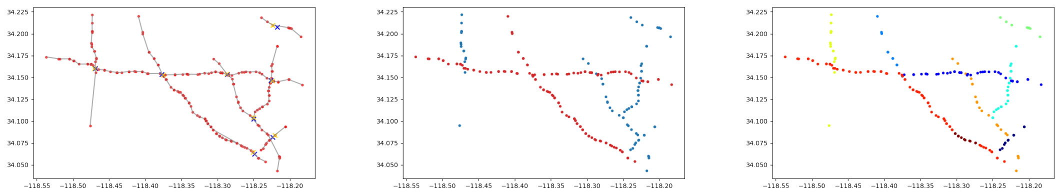
$$\hat{h}_{i,t} = \phi_h \left([A \, X \, W_h]^{\top}_{[i,:]} + [A \, (R_t \odot H_{t-1}) \, U_h]^{\top}_{[i,:]} + b_h \right) \tag{64}$$

$$h_{i,t} = z_t \odot h_{t-1} + (1 - z_t) \odot \hat{h}_t, \tag{65}$$

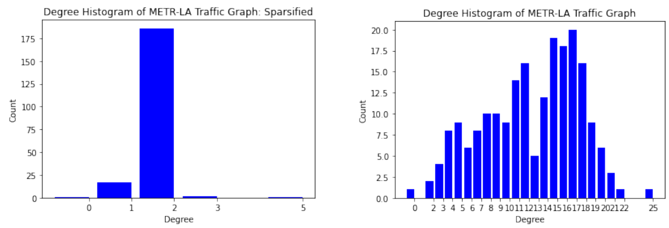
where $W_z, U_z, b_z, U_r, W_r, b_r, W_h, U_h, b_h$ are learnable weights and biases, σ_g is the sigmoid function, ϕ_g is the hyperbolic tangent, and $h_{i,0} = 0$ for all i at initialization. The crucial different from a vanilla GRU lies in eqn (64) where graph convolution replaces matrix multiplication.

We then modify the graph convolution in (64) from global weight sharing to tying weights among clusters of nodes, similar to the implementation in Appendix E.1 for Relax-$\mathcal{S}_{16}$. For example, in the case of two clusters (orbits), we change XW_h to

$$\texttt{swap} \left(\texttt{concat}[X_{c_1} W_{h,c_1}; X_{c_2} W_{h,c_2}] \right), \tag{66}$$



(a) Our faithful traffic graph (b) Graph clustering (2 clusters) (c) Graph clustering (9 clusters)



(d) Our faithful graph degree distribution (e) Original sensor graph degree distribution

Figure 5: METR-LA traffic graph: visualization, clustering, and degree distribution

where X_{c_i} denotes the submatrix of X including the rows of nodes from cluster i only, and W_{h,c_1}, W_{h,c_2} are two learnable matrices. In words, we perform cluster-specific linear transformation, combine the transformed features, and reorder the rows (i.e., `swap`) to ensure compatibility with the graph convolution.

Experiment Set-up. For our experiments, we use DCRNN model with 1 RNN layer and 1 diffusion step. We choose $T' = 3$ (i.e., 3 historical graph signals) and $T = 3$ (i.e., predict the next 3 period graph signals). We train all variants for 30 epochs using ADAM optimizer with learning rate 0.01. We report the test set performance selected by the best validation set performance.

E.2.1 Assumption Validation: Approximate Equivariant Map

Before applying our construction of approximate symmetries, we validate the assumption of the target function f^* being an approximately equivariant mapping using a trained DCRNN model as a proxy. We proceed as follows:

Data. We use the validation set of METR-LA (traffic graph signals in LA), which has 207 nodes and consists of $14,040$ input and output signals. Each input $X \in \mathbb{R}^{207 \times 2}$ represents the traffic volume and speed in the past at the 207 stations, and output $Y \in \mathbb{R}^{207}$ representing future traffic volume.

Model. We use a trained DCRNN model on our faithful graph, with input being 3 historical signals $\boldsymbol{X} = (X_{T-3}, X_{T-2}, X_{T-1}) \in \mathbb{R}^{3 \times 207 \times 2}$ to predict the future signals $\boldsymbol{Y} = (X_T, X_{T+1}, X_{T+2}) \in \mathbb{R}^{3 \times 207}$. We denote this model as f. It gives reasonable performance with Mean Absolute Error ≈ 3, and serves as a good proxy for the target (unknown) function f^*.

Neighbors. We take our faithful traffic graph that originally has 397 non-loop edges, and only consider a subset of 260 edges by thresholding the distance values to eliminate geometrically far-away nodes. This defines our 260 neighboring node pairs.

Equivariance error. For each node pair (i, j), we swap their input signals by interchanging the (i, j)-th slices in the node dimension of the tensor $\boldsymbol{X}$, denoted as $\boldsymbol{X}_{(i,j)}$, and check if the swapped output $\hat{\boldsymbol{Y}}_{(i,j)} = f(\boldsymbol{X}_{(i,j)})$ is close to the original output $\hat{\boldsymbol{Y}} = f(\boldsymbol{X})$ with (i, j)-th slices swapped. We measure "closeness" via the relative equivariant error at the node pair. Concretely, let $\boldsymbol{X}[i, j]$ denote the tensor slices at the (i, j) node pair, and $\boldsymbol{X}[j, i]$ being the swapped version by interchanging (i, j)-th slices. The relative different is computed as

$$\left|\hat{\boldsymbol{Y}}_{(i,j)}[j, i] - \hat{\boldsymbol{Y}}[i, j]\right| / \hat{\boldsymbol{Y}}[i, j],$$

where / denotes element-wise division. We then compute the mean relative equivariance error over all instances in the validation set, which equals to 5.17%. This gives concrete justification to enforce approximate equivariance in the traffic flow prediction problems.

References

[1] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[2] Fernando Gama, Joan Bruna, and Alejandro Ribeiro. Stability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 68:5680–5695, 2020.

[3] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

[4] Carlo Rovelli and Marcus Gaul. Loop quantum gravity and the meaning of diffeomorphism invariance. In *Towards Quantum Gravity: Proceeding of the XXXV International Winter School on Theoretical Physics Held in Polanica, Poland, 2–11 February 1999*, pages 277–324. Springer, 2000.

[5] Soledad Villar, David W Hogg, Weichi Yao, George A Kevrekidis, and Bernhard Schölkopf. The passive symmetries of machine learning. *arXiv preprint arXiv:2301.13724*, 2023.

[6] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[7] Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning under invariance and geometric stability, 2021.

[8] Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *International Conference on Machine Learning*, pages 2959–2969. PMLR, 2021.

[9] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*, 2021.

[10] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

[11] Bryn Elesedy. Group symmetry in pac learning. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.

[12] Bryn Elesedy. Provably strict generalisation benefit for invariance in kernel methods. *Advances in Neural Information Processing Systems*, 34:17273–17283, 2021.

[13] Arash Behboodi, Gabriele Cesa, and Taco Cohen. A pac-bayesian generalization bound for equivariant networks. *arXiv preprint arXiv:2210.13150*, 2022.

[14] Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33: 17605–17616, 2020.

[15] Jameson Cahill, Dustin G Mixon, and Hans Parshall. Lie pca: Density estimation for symmetric manifolds. *Applied and Computational Harmonic Analysis*, 65:279–295, 2023.

[16] Vasco Portilheiro. A tradeoff between universality of equivariant models and learnability of symmetries. *arXiv preprint arXiv:2210.09444*, 2022.

[17] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[18] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and deep locally connected networks on graphs. In *International Conference on Learning Representation*, 2014.

[19] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272, 2017.

[20] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[21] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[22] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, pages 3837–3845, 2016.

[23] F. Gama, A. G. Marques, G. Leus, and A. Ribeiro. Convolutional neural network architectures for signals supported on graphs. *IEEE Transactions on Signal Processing*, 67(4):1034–1049, 2019.

[24] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1): 97–109, 2019.

[25] Zhengdao Chen, Lei Chen, Soledad Villar, and Joan Bruna. Can graph neural networks count substructures? *Advances in neural information processing systems*, 33:10383–10395, 2020.

[26] Erik Henning Thiede, Wenda Zhou, and Risi Kondor. Autobahn: Automorphism-based graph neural nets, 2021. URL https://arxiv.org/abs/2103.01710.

[27] Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M Bronstein, and Haggai Maron. Equivariant subgraph aggregation networks. *arXiv preprint arXiv:2110.02910*, 2021.

[28] Giorgos Bouritsas, Fabrizio Frasca, Stefanos P Zafeiriou, and Michael Bronstein. Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[29] Fabrizio Frasca, Beatrice Bevilacqua, Michael M Bronstein, and Haggai Maron. Understanding and extending subgraph gnns by rethinking their symmetries. *arXiv preprint arXiv:2206.11140*, 2022.

[30] Bohang Zhang, Guhao Feng, Yiheng Du, Di He, and Liwei Wang. A complete expressiveness hierarchy for subgraph gnns via subgraph weisfeiler-lehman tests. *CoRR*, abs/2302.07090, 2023.

[31] Jianfei Gao, Yangze Zhou, and Bruno Ribeiro. Double permutation equivariance for knowledge graph completion. *arXiv preprint arXiv:2302.01313*, 2023.

[32] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

[33] Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI Conference on Artificial Intelligence*, pages 4602–4609, 2019.

[34] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

[35] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[36] Yashil Sukurdeep et al. *Elastic shape analysis of geometric objects with complex structures and partial correspondences*. PhD thesis, Johns Hopkins University, 2023.

[37] Jameson Cahill, Joseph W Iverson, Dustin G Mixon, and Daniel Packer. Group-invariant max filtering. *arXiv preprint arXiv:2205.14039*, 2022.

[38] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv:1802.08219*, 2018.

[39] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Neural Information Processing Systems*, 33, 2020.

[40] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

[41] Soledad Villar, David W Hogg, Kate Storey-Fisher, Weichi Yao, and Ben Blum-Smith. Scalars are universal: Equivariant machine learning, structured like classical physics. *Advances in Neural Information Processing Systems*, 34:28848–28863, 2021.

[42] Ben Finkelshtein, Chaim Baskin, Haggai Maron, and Nadav Dym. A simple and universal rotation equivariant point-cloud network. In *Topological, Algebraic and Geometric Learning Workshops 2022*, pages 107–115. PMLR, 2022.

[43] Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. *Advances in Neural Information Processing Systems*, 31:10117–10126, 2018.

[44] Maurice Weiler, Patrick Forré, Erik Verlinde, and Max Welling. Coordinate independent convolutional networks–isometry and gauge equivariant convolutions on riemannian manifolds. *arXiv preprint arXiv:2106.06020*, 2021.

[45] Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23078–23091. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22aa.html.

[46] Soledad Villar, Weichi Yao, David W Hogg, Ben Blum-Smith, and Bianca Dumitrascu. Dimensionless machine learning: Imposing exact units equivariance. *Journal of Machine Learning Research*, 24(109):1–32, 2023.

[47] Taco S. Cohen, Mario Geiger, Jonas Koehler, and Max Welling. Spherical cnns, 2018.

[48] Jiaqi Han, Wenbing Huang, Tingyang Xu, and Yu Rong. Equivariant graph hierarchy-based neural networks. *Advances in Neural Information Processing Systems*, 35:9176–9187, 2022.

[49] Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3419–3430. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/garg20c.html.

[50] Pascal Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. *Advances in Neural Information Processing Systems*, 34:27043–27056, 2021.

[51] Christopher Morris, Floris Geerts, Jan Tönshoff, and Martin Grohe. Wl meet vc, 2023.

[52] Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. In *Advances in Neural Information Processing Systems*, 2022.

[53] Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22 (272):1–59, 2021.

[54] Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. *Advances in Neural Information Processing Systems*, 33: 1702–1712, 2020.

[55] Beatrice Bevilacqua, Yangze Zhou, and Bruno Ribeiro. Size-invariant graph representations for graph classification extrapolations. In *International Conference on Machine Learning*, pages 837–851. PMLR, 2021.

[56] Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*, 2020.

[57] Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual pathway priors for soft equivariance constraints. *Advances in Neural Information Processing Systems*, 34:30037–30049, 2021.

[58] Alexander Bogatskiy, Sanmay Ganguly, Thomas Kipf, Risi Kondor, David W Miller, Daniel Murnane, Jan T Offermann, Mariel Pettee, Phiala Shanahan, Chase Shimmin, et al. Symmetry group equivariant architectures for physics. *arXiv preprint arXiv:2203.06153*, 2022.

[59] Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. doi: https://doi.org/10.1002/cpa.21413.

[60] William Fulton and Joe Harris. *Representation theory: a first course*, volume 129. Springer Science & Business Media, 2013.

[61] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. In *International conference on machine learning*, pages 2892–2901. PMLR, 2017.

[62] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

[63] Christian Borgs, Jennifer T Chayes, László Lovász, Vera T Sós, and Katalin Vesztergombi. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

[64] László Lovász and Balázs Szegedy. Szemerédi's lemma for the analyst. *GAFA Geometric And Functional Analysis*, 17(1):252–270, 2007.

[65] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[66] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3425–3435, 2019.

[67] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[68] Bruce Sagan. *The symmetric group: representations, combinatorial algorithms, and symmetric functions*, volume 203. Springer Science & Business Media, 2001.

[69] Jean-Pierre Serre et al. *Linear representations of finite groups*, volume 42. Springer, 1977.

[70] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[71] Ningyuan Teresa, David W. Hogg, and Soledad Villar. Dimensionality reduction, regularization, and generalization in overparameterized regressions. *SIAM Journal on Mathematics of Data Science*, 4(1):126–152, feb 2022. doi: 10.1137/20m1387821. URL `https://doi.org/10.1137%2F20m1387821`.

[72] Somesh Das Gupta. Some aspects of discrimination function coefficients. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 387–400, 1968.

[73] Ron Levie. A graphon-signal analysis of graph neural networks. 2023.

[74] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019.

[75] Hosagrahar V Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.