Improved Generalization Bounds of Group Invariant / Equivariant Deep Networks via Quotient Feature Spaces

Akiyoshi Sannai[1,2] **Masaaki Imaizumi**[3,1] **Makoto Kawano**[3]

[1]RIKEN Center for Advanced Intelligence Project, Chuo, Tokyo, Japan
[2]Keio University, Minato, Tokyo, Japan
[3] The University of Tokyo, Bunkyo, Tokyo, Japan

Abstract

Numerous invariant (or equivariant) neural networks have succeeded in handling invariant data such as point clouds and graphs. However, a generalization theory for the neural networks has not been well developed, because several essential factors for the theory, such as network size and margin distribution, are not deeply connected to the invariance and equivariance. In this study, we develop a novel generalization error bound for invariant and equivariant deep neural networks. To describe the effect of invariance and equivariance on generalization, we develop a notion of a *quotient feature space*, which measures the effect of group actions for the properties. Our main result proves that the volume of quotient feature spaces can describe the generalization error. Furthermore, the bound shows that the invariance and equivariance significantly improve the leading term of the bound. We apply our result to specific invariant and equivariant networks, such as DeepSets [Zaheer et al., 2017], and show that their generalization bound is considerably improved by $\sqrt{n!}$, where $n!$ is the number of permutations. We also discuss the expressive power of invariant DNNs and show that they can achieve an optimal approximation rate. Our experimental result supports our theoretical claims.

1 INTRODUCTION

Group invariant (or equivariant) deep neural networks have been extensively utilized in data analysis [Shawe-Taylor, 1989, 1993, Ntampaka et al., 2016, Ravanbakhsh et al., 2016, Faber et al., 2016, Cohen and Welling, 2016, Zaheer et al., 2017, Li et al., 2018a, Su et al., 2018, Li et al., 2018b, Yang et al., 2018, Xu et al., 2018, Lenssen et al., 2018, Cohen et al., 2019]. A typical example is permutation invariant deep neural networks for point cloud data. The data are given as a set of points, and permuting points in the data does not change the result of its prediction [Zaheer et al., 2017, Li et al., 2018a, Su et al., 2018, Li et al., 2018b, Yang et al., 2018, Xu et al., 2018, Ntampaka et al., 2016, Ravanbakhsh et al., 2016, Faber et al., 2016]. Another example is graph neural networks for graph data, that are represented by a column-and-row-permutation invariant adjacency matrix [Bruna et al., 2013, Henaff et al., 2015, Monti et al., 2017, Ying et al., 2018]. The group invariant and equivariant neural networks can significantly improve the accuracy of prediction with limited data size and network size [Zaheer et al., 2017, Li et al., 2018b,a]. Their theoretical properties have been investigated as well. Universal approximation properties are proved for several invariant and equivariant neural networks [Yarotsky, 2018, Maron et al., 2019, Sannai et al., 2019, Segol and Lipman, 2019, Ravanbakhsh, 2020].

Despite the impact and high empirical accuracy, the generalization error of group invariant / equivariant neural networks has not been well clarified yet. This is because there are several theoretical difficulties in connecting invariance with generalization theory. First, invariance is not strongly connected to common factors that are important to the theory. The generalization error bounds of ordinary deep neural networks are mainly controlled by their depth, width, number of trainable parameters, and margin distributions [Anthony and Bartlett, 2009, Neyshabur et al., 2015, Bartlett et al., 2017]. However, invariance and equivariance are determined independently of these factors. Second, there are few quantitative features which can assess invariance and equivariance. Without a quantitative criterion, it is not possible to measure how invariance and equivariance affect on generalization errors.

In this study, we establish a unified generalized error bound by developing a quantitative measure to describe the effects of invariance and equivariance. For a deep neural network f, let $R(f)$ be its expected loss and $R_m(f)$ be its empirical loss with m training samples. For a set of neural networks

$\mathcal{F}$, we are interested in the following value

$$\mathcal{G}(\mathcal{F}) := \sup_{f \in \mathcal{F}} |R(f) - R_m(f)|, \quad (1)$$

which is referred as a bound on *generalization gap* or *generalization error*. Our theory can describe significant improvements in the generalization error bounds of invariant / equivariant neural networks. We summarize our results as follows.

(i) Generalization Bound with Quotient Feature Space: We develop a notion of a *quotient feature space* (QFS) and prove that the generalization error bound of invariant / equivariant neural networks is described by the volume of QFSs. For a finite group G, we define a quotient feature map $\phi_G : \mathbb{R}^n \to \mathbb{R}^n/G$ and then define a QFS as $\Delta_G := \phi_G([0,1]^n)$, which is regarded as a feature space associated with G. Our results show that the generalization error is proportional to the square root of the volume of Δ_G (invariant case) or $\Delta_{\mathrm{St}(G)}$ (equivariant case), where $\mathrm{St}(G) \subset G$ is a subgroup of elements whose first coordinates are fixed, named a *stabilizer subgroup*. In short, with a set of G-invariant deep neural networks $\mathcal{F}^G$, we obtain the following intuition:

$$\mathcal{G}(\mathcal{F}^G) \propto \sqrt{\mathrm{vol}(\Delta_G)}$$

Theorem 2 shows a rigorous statement. Figure 2 provides examples of QFSs with several G.

(ii) Roles of Invariance for Generalization: We identify how invariance improves generalization through the result with QFSs. We consider the symmetric group $G = S_n$ for example. In this case, we derive the following bound:

Theorem 1 (Informal Corollary 1)**.** *Let $\mathcal{F}^{S_n}$ be a set of S_n-invariant deep neural networks. For any $\varepsilon > 0$,*

$$\mathcal{G}(\mathcal{F}^{S_n}) \leq O\left(\sqrt{\frac{1}{n!\, m^{2/n}}}\right) + \sqrt{\frac{2\log(1/2\varepsilon)}{m}},$$

holds with at least probability $1 - 2\varepsilon$.

This bound reveals two properties of invariant networks. First, the scale of the bound is improved by $\sqrt{n!}$. This result follows the fact that the generalization gap is proportional to the size of QFSs. This improvement is significant, since n takes a large value in recent point cloud data, for example $n > 1,000$. Second, it slows down the convergence rate in a number of samples m. This deterioration is a price of gaining the factorial improvement in n. However, as shown in Figure 1, the factorial improvement greatly outweighs the rate deterioration.

We have mainly two technical contributions. First, we define the notion of a QFS and show its geometric properties, then derive its volume with a wide class of G. Second, we show a connection between a set of invariant / equivariant neural networks and the volume of QFSs, then describe their generalization bounds by the volume. Furthermore, we investigate the expressive power of S_n-invariant deep neural networks and show their expressive power attains an optimal rate.

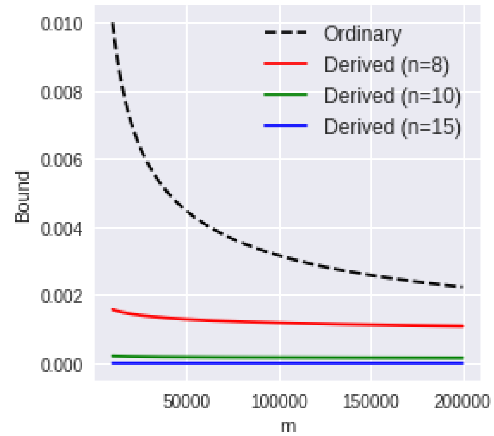

Figure 1: Order of the bound for the generalization gap against m. *Ordinary* (dashed line) denotes $(1/\sqrt{m})$ without invariance, and *Derived* (colored lines) denote the bound $(1/\sqrt{n! m^{2/n}})$ with $n \in \{8, 10, 15\}$. Regardless of the effect of m, the derived bound gets tight sharply as n increases.

1.1 RELATED WORK

There are several works studying the generalized performance and sample complexity of neural networks with invariance / equivariance. Shawetaylor [1995] shows that the sample complexity increases by a number of equivalent classes. The closest work with our study is Sokolic et al. [2017], which considers a general algorithm for the classification problem. Their generalization bound is proportional to $\sqrt{1/T}$, where T is the number of *transformations* generated by the invariance property. While their research is excellent, we improve their work in two ways. (I) Our result has a more concrete structure: our generalization bound describes an explicit role of invariance and equivariance through the notion of QFSs. Owing to QFSs, our result can be applied to various cases such as graphs. (II) We relax their strong assumptions on stability and provide accurate analysis. We provide its detail in Section 7.2. In fact, our theoretical results are not limited to deep neural networks. However, most of the models that can control invariant data with large n, such as point clouds and large graphs, are mainly handled by deep neural networks [Zaheer et al., 2017, Maron et al., 2018, 2019, 2020]. Hence, we regard neural networks as the main application of our theory.

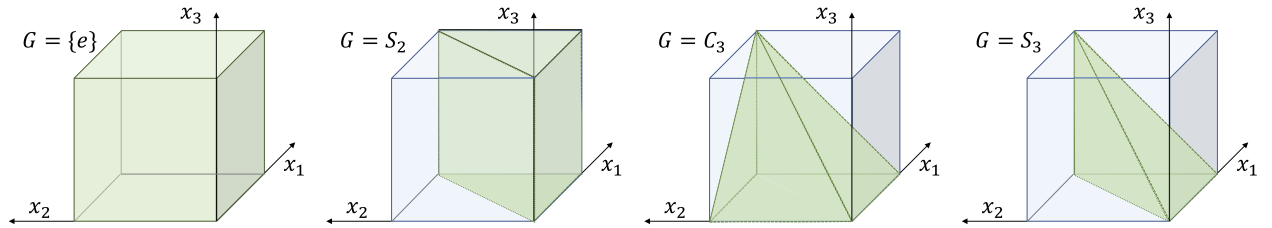

Figure 2: Example of quotient feature spaces with $n = 3$. (i) trivial group case ($G = \{e\}$): $\Delta_{\{e\}} = [0, 1]^3$. (ii) symmetric group case ($G = S_2$): $\Delta_{S_2} = \{x \in [0, 1]^3 \mid x_1 \geq x_2\}$. (iii) cyclic group case ($G = C_3$): $\Delta_{C_3} = \{x \in [0, 1]^3 \mid x_1 \geq x_2 \geq x_3\} \cup \{x \in [0, 1]^3 \mid x_1 \leq x_2 \leq x_3\}$. (iv) symmetric group case ($G = S_3$): $\Delta_{S_3} = \{x \in [0, 1]^3 \mid x_1 \geq x_2 \geq x_3\}$.

Deep Network	Group	$\mathrm{vol}(\Delta_G)$	$\mathrm{vol}(\Delta_{\mathrm{St}(G)})$
Deep Sets [Zaheer et al., 2017]	S_n	$O(1/(n!))$	$O(1/((n-1)!))$
G-CNN [Cohen and Welling, 2016]	C_4	$O(1/4)$	$O(1)$
Graph Network [Maron et al., 2018]	$S_n \subset S_{n^2}$	$O(1/(\#\text{of nodes})!)$	$O(1/((\#\text{of nodes}) - 1)!)$
Tensor Network [Maron et al., 2019]	$G \subset S_n$	$O(1/\|G\|)$	$O(1/\|\mathrm{St}(G)\|)$
DSS [Maron et al., 2020]	$S_n \times G'$ $(G' \subset S_N)$	$O(1/(n!\|G'\|))$	$O(1/((n-1)!\|\mathrm{St}(G')\|)$

Table 1: Examples of invariant / equivariant DNNs utilized in practice. S_n denotes a symmetric group of order $n!$, and C_n denotes a cyclic group of order n. G denotes a subgroup of the permutation group S_n of axes of the input space $\mathbb{R}^n$. G' denotes a subgroup of the permutation group S_N of axes of the input space $\mathbb{R}^{n\times N}$. We set $\mathrm{vol}(\Delta_G) = \mathcal{N}_{\varepsilon,\infty}(\Delta_G)$, where $\mathcal{N}_{\varepsilon,\infty}(\Delta_G)$ is a covering number of Δ_G in terms of $\|\cdot\|_\infty$. DSS was referred to as "Deep Sets for Symmetric elements layers" [Maron et al., 2020].

1.2 NOTATION

For a vector $b \in \mathbb{R}^D$, its d-th element is denoted by b_d. For a function $f : \Omega \to \mathbb{R}$ with a set Ω, $\|f\|_{L^q} := (\int_\Omega |f(x)|^q dx)^{1/q}$ denotes the L^q-norm for $q \in [0, \infty]$. For a subset $\Lambda \subset \Omega$, $f_{\upharpoonright\Lambda}$ denotes a restriction of f to Λ. $C(\Omega)$ denotes a set of continuous functions on Ω. For an integer z, $z! = \prod_{j=1}^n j$ denotes a factorial of z. For a set Ω, id_Ω or id denotes the identity map on Ω, namely $\mathrm{id}_\Omega(x) = x$ for any $x \in \Omega$. For a subset $\Delta \subset \mathbb{R}^n$, $\mathrm{int}(\Delta)$ denotes a set of inner points of a set Δ. For metric spaces Δ and Δ', $\Delta \cong \Delta'$ denotes they are isomorphic as metric spaces. The supplementary material maintains all full proofs.

2 DEFINITION AND PROBLEM SETTING

2.1 INVARIANCE / EQUIVARIANCE AND DEEP NEURAL NETWORK

We provide a general concept of the invariance and equivariance of functions. Throughout this paper, we consider a finite group $G \leq S_n$, where S_n denotes the symmetric group.

Definition 1 (Invariant / Equivariant Function)**.** For a group G acting on $\mathbb{R}^n$ and $\mathbb{R}^M$, a function $f\colon \mathbb{R}^n \to \mathbb{R}^M$ is

- *G-invariant* if $f(g \cdot x) = f(x)$ holds for any $g \in G$ and any $x \in \mathbb{R}^n$,
- *G-equivariant* if $f(g \cdot x) = g \cdot f(x)$ holds for any $g \in G$ and any $x \in \mathbb{R}^n$.

For a set Ω, $C^G(\Omega)$ denotes a set of G-invariant an continuous functions on Ω.

We formulate deep neural networks (DNNs) with invariance and equivariance. In this study, we consider fully connected DNNs with the ReLU activation function $\mathrm{ReLU}(x) = \max(0, x)$. Let us consider a layer-wise map $Z_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$ defined by $Z_i(x) = \mathrm{ReLU}(W_i x + b_i)$, where $W_i \in \mathbb{R}^{d_{i+1}\times d_i}$ and $b_i \in \mathbb{R}^{d_{i+1}}$ for $i = 1, ..., H$. Here, H is a depth, and d_i is a width of the i-th layer. Then, a function by DNNs has the following formulation

$$f(x) := Z_H \circ Z_{H-1} \ldots Z_2 \circ Z_1(x). \quad (2)$$

Further, let $\mathcal{F}$ be a set of functions with the form (2).

We define a function by invariant and equivariant DNNs.

Definition 2 (Invariant / Equivariant Deep Neural Network)**.** A function $f \in \mathcal{F}$ is a *G-invariant / equivariant DNN*, if f is a G-invariant / equivariant function.

This definition is a general notion and represents several explicit invariant DNNs. We provide several representative examples as follows.

Example 1 (Deep Sets)**.** A permutation-invariant (S_n-invariant) DNN was developed by Zaheer et al. [2017].

Its architecture has J middle permutation-equivariant (S_n-equivariant) layers $Z_1, ..., Z_J$, a permutation-invariant linear layer Z_L, and a fully-connected layer Z_F. Each equivariant layer maintains a parameter matrix $W_i = \lambda \boldsymbol{I} + \gamma(\mathbf{1}\mathbf{1}^\top),\ \lambda, \gamma \in \mathbb{R}, \mathbf{1} = [1, ..., 1]^\top$, which makes Z_j be equivariant. Then, a DNN $f = Z_F \circ Z_L \circ Z_J \circ \cdots Z_1$ is a permutation-invariant DNN.

Example 2 (Tensor Network)**.** For a finite group $G \subset S_n$, a G-invariant / equivariant DNN was developed by Maron et al. [2019] using a notion of higher-order tensors. The study considered a tensor $W \in \mathbb{R}^{n^k \times a}$ and an action $g \in G$ on the tensor as $(g \cdot W)_{i_1,..,i_k,j} = W_{g^{-1}(i_1),..,g^{-1}(i_k),j}$, for $i_{k'} = 1, ..., n, k' = 1, ..., k$, and $j = 1, ..., a$. With the action, the study developed a G-invariant / equivariant DNN. Since G is a finite group, the model is a specific case of our setting.

2.2 FORMULATION OF LEARNING PROBLEM

We formulate our learning problem with DNNs. Let $I = [0, 1]^n$ be an input space with $n \in \mathbb{N}$ and $\mathbb{R}^M$ be an output space with $M \in \mathbb{N}$. Let $L : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ be a loss function which satisfies $\sup_{y,y' \in \mathcal{Y}} |L(y, y')| \leq 1$ and 1-Lipschitz continuous. Let $P^*(x, y)$ be a distribution on $I \times \mathbb{R}^M$ which generate data, and $R(f) = \mathbb{E}_{(X,Y)\sim P^*}[L(f(X), Y)]$ for $f : I \to \mathbb{R}^M$ be the expected loss of f. Suppose we have a training dataset $\{(X_1, Y_1), ..., (X_m, Y_m)\}$ of size m which is independently generated from P^*. Let $R_m(f) := m^{-1} \sum_{i=1}^m L(f(X_i), Y_i)$ be an empirical loss with f. Our interest is to bound $\mathcal{G}(\mathcal{F})$ as (1) which illustrates how minimizing $R_m(f)$ on $\mathcal{F}$ affects $R(f)$.

3 QUOTIENT FEATURE SPACES

We provide a notion of a *quotient feature space* (QFS), which is a key factor in connecting invariance / equivariance and generalization. With a quotient space $\mathbb{R}^n/G$ with G, we consider a map

$$\phi_G : \mathbb{R}^n \to \mathbb{R}^n/G \text{ such as } \phi_G(x) = \{g \cdot x \mid \forall g \in G\},$$

named a *quotient map*. By the definition of g, such ϕ_G always exists. With this notion, we define a QFS.

Definition 3 (Quotient Feature Space)**.** For a finite group G, a *quotient feature space* is defined as

$$\Delta_G := \phi_G(I).$$

We can regard a QFS as a feature space with G. We prove that a QFS can equip a distance if $g \in G$ preserves a distance in I, which is a fundamental property of feature spaces.

Proposition 1 (Distance on QFS)**.** *For a finite group G, we define a function $d_G : \mathbb{R}^n/G \times \mathbb{R}^n/G \to \mathbb{R}_{\geq 0}$ as*

$$d_G(y, y') = \inf\{\|x - x'\|_2 \mid \phi_G(x) = y, \phi_G(x') = y'\}.$$

Then, d_G is a distance on $\mathbb{R}^n/G$.

Intuitively, the distance d_G for $\mathbb{R}^n/G$ is an infimum of a sum of pairwise distances of points $\{x \mid \phi_G(x) = y\}$ and $\{x' \mid \phi_G(x') = y'\}$. $g \in G$ maintains the distance when G is a finite group. We also remark that this proposition does not hold for some infinite groups.

3.1 VOLUME MEASUREMENT OF QFS

We measure volume of Δ_G, which is a critical factor for a generalization bound of invariant / equivariant DNNs. We consider two cases: (i) the symmetric group $G = S_n$, and (ii) a finite group G. We measure the volume of a set Ω using a *covering number* $\mathcal{N}_{\varepsilon,\infty}(\Omega) := \inf\{N \mid \exists \{x_j\}_{j=1}^N \subset \Omega, \text{ s.t. } \cup_{j=1}^N \{x \mid \|x - x_j\|_\infty \leq \varepsilon\} \supset \Omega\}$.

3.1.1 Symmetric Group Case

We begin with the symmetric group $G = S_n$. It is convenient to study S_n as a first step, because we can derive an explicit formulation of ϕ_{S_n} and Δ_{S_n}. With the case, an action $\sigma \in S_n$ is a permutation of indexes of $x = (x_1, ..., x_n) \in I$. For $i = 1, ..., n$, we define a map $\max_i(\{x_1, ..., x_n\})$ which returns the i-th largest element of $\{x_1, ..., x_n\}$.

Proposition 2 (QFS of S_n)**.** *Define a set $\Delta \subset I$ as $\Delta := \{x \in I \mid x_1 \geq x_2 \geq \cdots \geq x_n\}$, and a map $\phi : \mathbb{R}^n \to \mathbb{R}^n$ as $\phi(x) := (\max_1(\{x_1, .., x_n\}), ..., \max_n(\{x_1, .., x_n\}))$. Then, we obtain $\phi(\Delta) \cong \Delta_{S_n}$.*

Figure 2 illustrates Δ_{S_n} for some n. Intuitively, any element of I corresponds to some element of Δ_{S_n} with an existing action $\sigma \in S_n$, namely, $I = \cup_{\sigma \in S_n} \{\sigma \cdot x \mid x \in \Delta\}$ holds. With the help of the explicit formulation of Δ_{S_n}, we can measure its size. Since $\Delta_{S_n} \subset I$ holds, we can measure its volume by the Euclidean distance as follows:

Lemma 1 (Volume of Δ_{S_n})**.** *There is a constant C such that for small enough $\varepsilon > 0$, we obtain*

$$\mathcal{N}_{\varepsilon,\infty}(\Delta_{S_n}) \leq C/(n!\, \varepsilon^n).$$

Lemma 1 provides an important claim: the volume of Δ_{S_n} is proportional to $1/(n!)$, i.e., the volume significantly decreases with the increases in n. The term ε^{-n} is usual for covering numbers, i.e., $\mathcal{N}_{\varepsilon,\infty}(I) \leq C/\varepsilon^n$ holds, hence the factorial improvement by $1/(n!)$ comes from S_n-invariance.

3.1.2 General Finite Group Case

We consider a general finite group G and its corresponding QFS, by studying Δ_G and measuring its covering volume. We first prepare several notions. For a group G, $|G|$ denotes its number of elements, named an *order* of G. For a subgroup $H \subset G$, a set $\{g_1, .., g_K | g_k \in G\}$ is defined as a *complete system of representatives of* $H \backslash G$ if $K = |G|/|H|$ and $G = \cup_{k=1}^{K} H \cdot g_k$ hold. For any G and H, we can always find the complete system. Also, we define $\Delta_k := g_k \cdot \Delta_{S_n}$. Then, we achieve the following result:

Proposition 3. *Let $\{g_1, .., g_K | g_k \in S_n, k = 1, ..., K\}$ be a complete system of representatives of $G \backslash S_n$. Then, $\Delta_k \cong \Delta_{S_n}$ holds as metric spaces for all $k = 1, ..., K$. Furthermore, its induced set $\widetilde{\Delta}_G := \bigcup_{k=1}^{|S_n|/|G|} \Delta_k$ satisfies $\phi_G(\widetilde{\Delta}_G) = \Delta_G$.*

Proposition 3 shows that we can describe Δ_G by $\widetilde{\Delta}_G$ which is a combination of complete systems of representatives of $G \backslash S_n$. Intuitively, we can define $\widetilde{\Delta}_G$ by a union of several transformed Δ_{S_n}.

We describe an example with $n = 3$ and $G = S_2$. A complete system of representatives of $S_2 \backslash S_3$ can be $\{g_1, g_2, g_3\} \subset S_3$ such that g_1 is an identity, g_2 is a transposition of the 2nd and 3rd elements, and g_3 is a cyclic permutation. In other words, we have $g_3 \cdot 1 = 2, g_3 \cdot 2 = 3$, and $g_3 \cdot 3 = 1$. Moreover, we have $\widetilde{\Delta}_{S_2} = \Delta_{S_2}$. Then, we can represent Δ_{S_2} by Δ_{g_k} with $k = 1, 2, 3$ as $\Delta_{S_2} = g_1 \cdot \Delta_{S_3} \cup g_2 \cdot \Delta_{S_3} \cup g_3 \cdot \Delta_{S_3}$. According to Figure 3, Δ_{S_2} is a union of $\Delta_{S_3} (= g_1 \cdot \Delta_{S_3})$, reflected Δ_{S_3} $(= g_2 \cdot \Delta_{S_3})$, and rotated Δ_{S_3} $(= g_3 \cdot \Delta_{S_3})$.

We can evaluate the volume of Δ_G by $|G|$ as follows:

Lemma 2. *There exists a constant $C > 0$, such that for small enough $\varepsilon > 0$, we obtain*

$$\mathcal{N}_{\varepsilon,\infty}(\Delta_G) \leq C/(|G|\, \varepsilon^n).$$

Similar to Lemma 1, the result of Lemma 2 states that the covering volume of Δ_G is improved by $|G|$. Since $|S_n| = n!$ holds, Lemma 2 is a generalization of Lemma 1. Table 1 contains examples of G.

3.2 COVERING NUMBERS OF QFS

We show several technical inequalities to present a relationship between G-invariant DNNs and Δ_G. Namely, we show that a covering number of a set $\mathcal{F}^G(I) = \{f : I \to \mathbb{R}^M \mid f \text{ is a } G\text{-invariant DNN.}\}$ is evaluated by comparison with the volume of Δ_G. We also define $\mathcal{F}(\Delta_G) := \{f : \Delta_G \to \mathbb{R}^M \mid f \text{ is a DNN.}\}$. We note that $f \in \mathcal{F}(\Delta_G)$ is an ordinary DNN rather than a G-invariant DNN.

First, we derive a corresponding map between the two functional sets.

Proposition 4. *ϕ_G induces a bijection $\widehat{\phi}_G : C(\Delta_G) \to C^G(I)$. Further, $f \in C(\Delta_G)$ is K-Lipschitz continuous if and only if $\widehat{\phi}_G(f)$ is K-Lipschitz continuous.*

Using the corresponding map, we evaluate the volume of $\mathcal{F}^G(I)$ by $\mathcal{F}(\Delta_G)$. The following result presents the claim.

Proposition 5. *For any $\varepsilon > 0$, we obtain*

$$\mathcal{N}_{\varepsilon,\infty}(\mathcal{F}^G(I)) \leq \mathcal{N}_{\varepsilon,\infty}(\mathcal{F}(\Delta_G)).$$

This inequality shows that the set of G-invariant DNNs on I is bounded by the volume of the set of DNNs on Δ_G *without* invariance.

Finally, we evaluate the volume of $\mathcal{F}(\Delta_G)$ in terms of the volume of Δ_G. We provide an inequality which bounds the covering number of $\mathcal{F}(\Delta_G)$ by a *polynomial* of the volume of Δ_G, whereas the commonly used inequality only includes the *logarithm* of the volume of Δ_G, such as the result in Section 10.2 in Anthony and Bartlett [2009].

Proposition 6. *Suppose that any function in $\mathcal{F}(\Delta_G)$ is C_Δ-Lipschitz continuous and uniformly bounded by B with constants $C_\Delta, B > 0$. Then, with an existing constant $c > 0$ and C in Lemma 1, for any $\delta > 0$, we obtain*

$$\log \mathcal{N}_{2C_\Delta \delta,\infty}(\mathcal{F}(\Delta_G)) \leq \mathcal{N}_{\delta,\infty}(\Delta_{S_n}) \log(8c^2 B/\delta).$$

Combining this result with Proposition 5, we can utilize $\mathcal{N}_{\varepsilon,\infty}(\Delta_{S_n})$ as the quantitative measure to evaluate the volume of the set of G-invariant DNNs.

Remark 1 (Linear bound in $\mathcal{N}_{\delta,\infty}(\Delta_{S_n})$)**.** In Proposition 6, it is important to note that logarithm of $\mathcal{N}_{2C_\Delta \delta,\infty}(\mathcal{F}(\Delta_G))$ is linearly bounded by $\mathcal{N}_{\delta,\infty}(\Delta_{S_n})$. In general, log of $\mathcal{N}_{2C_\Delta \delta,\infty}(\mathcal{F}(\Delta_G),)$ is bounded by a number of parameters of DNNs [Anthony and Bartlett, 2009] or parameter norms [Bartlett et al., 2017]. However, these values have little to do with invariance and therefore cannot give tight bounds. We instead consider the volume of Δ_{S_n} as a value related to invariance and achieve the linear bound in $\mathcal{N}_{\delta,\infty}(\Delta_{S_n})$.

4 GENERALIZATION BOUND FOR INVARIANT DNNS

We derive a generalization bound with QFSs and show that invariance can effectively improve the generalization performance of DNNs. Utilizing the results above, we have the following main result:

Theorem 2 (Generalization of Invariant DNN)**.** *Suppose any $f \in \mathcal{F}^G = \mathcal{F}^G(I)$ is uniformly bounded by 1. Then, for any $\varepsilon > 0$, there exist a constant $C > 0$ that are independent of n, m and ε, and the following inequality holds*

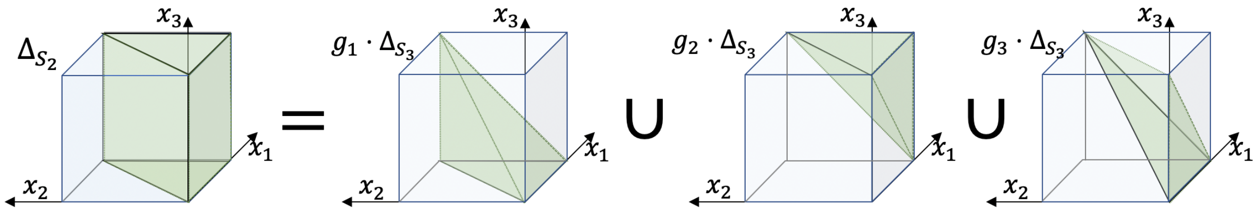

Figure 3: Illustration of $\Delta_{S_2} = \widetilde{\Delta}_{S_2} = g_1 \cdot \Delta_{S_3} \cup g_2 \cdot \Delta_{S_3} \cup g_3 \cdot \Delta_{S_3}$. The blue cube is I, and the green polyhedrons are Δ_{S_2} and $g_k \cdot \Delta_{S_3}, k = 1, 2, 3$.

with probability at least $1 - 2\varepsilon$*:*

$$\mathcal{G}(\mathcal{F}^G) \leq \underbrace{\sqrt{\frac{C}{|G|\, m^{2/n}}}}_{=:I_1} + \underbrace{\sqrt{\frac{2\log(1/2\varepsilon)}{m}}}_{=:I_2}.$$

The main term I_1 of the bound is interpreted to maintain the relation $I_1 \propto \sqrt{\mathcal{N}_{\varepsilon,\infty}(\Delta_G)}$, hence the volume of QFSs describes the effect of invariance on the generalization error. Obviously, I_1 is improved as $\sqrt{|G|}$ increases. Although the convergence rate of the main term in m gets slow as n increases, an increase in $\sqrt{|G|}$ reduces the error, as shown in the following specific example. Here, we note that we can regard I_2 as a relatively negligible term.

With the case $G = S_n$, the result in Theorem 2 yields a more explicit bound:

Corollary 1 (Generalization of S_n-invariant DNN)**.** *Consider the same setting as Theorem 2. Then, for any* $\varepsilon > 0$*, there exists a constant* $C > 0$ *and the following inequality holds with probability at least* $1 - 2\varepsilon$*:*

$$\mathcal{G}(\mathcal{F}^{S_n}) \leq \sqrt{\frac{C}{n!\, m^{2/n}}} + \sqrt{\frac{2\log(1/2\varepsilon)}{m}}.$$

Corollary 1 follows the order $|S_n| = n!$. Since n is large in practice, e.g., a number of points in point cloud data or a number of nodes for graph data, the term $n!$ significantly improves the bound.

Remark 2 (Convergence rate in m)**.** In the result, the convergence rate to the sample size m slows down as n increases, but the improvement of the bound with increasing n more than cancels this out. Since a decay by the factorial term $n!$ is sufficiently faster than any polynomial convergence in n. For a practical example with an experiment ($m = 9843, n = 100$) by Zaheer et al. [2017] with the ModelNet40 dataset [Wu et al., 2015], our bound $O(1/\sqrt{n!m^{2/n}}) \approx O(10^{-156})$ is *significantly tighter* than an ordinary bound $O(1/\sqrt{m}) \approx O(10^{-1.99})$.

Proof sketch for Theorem 2: We prove Theorem 2 by the following three steps.

First, we apply the well-known Rademacher complexity bound (e.g., Lemma A.5 in Bartlett et al. [2017]) and obtain the following inequality with probability at least $1 - 2\varepsilon$

$$\mathcal{G}(\mathcal{F}^G) \leq \sqrt{\frac{2\log(1/2\varepsilon)}{m}} + \inf_{\alpha \geq 0}\left\{4\alpha + \frac{12}{\sqrt{m}}\int_{\alpha}^{\sqrt{m}} \sqrt{2\log 2\mathcal{N}_{\delta,\infty}(\mathcal{F}^G(I))}d\delta\right\}. \tag{3}$$

Second, we bound the term $\log \mathcal{N}_{\delta,\infty}(\mathcal{F}^G(I))$ in (3) by $\log \mathcal{N}_{\delta,\infty}(\mathcal{F}(\Delta_G))$ by using the result in Proposition 4. This enables us to evaluate the error with G-invariance using Δ_G.

Third, we bound $\log \mathcal{N}_{\delta,\infty}(\mathcal{F}(\Delta_G))$ by the term with $\mathcal{N}_{\delta,\infty}(\Delta_G)$. To achieve this bound for bounding the volume of functional sets by that of its domain, we provide Proposition 6 in the supplementary material. Then, we combine Lemma 2 and get the statement of Theorem 2. □

5 GENERALIZATION BOUND FOR EQUIVARIANT DNNS

We derive a generalization bound for equivariant DNNs. To this aim, we require a covering number of the following set $\widetilde{\mathcal{F}}^G(I) = \{\widetilde{f} : I \to \mathbb{R}^n \mid \widetilde{f}$ is a G-equivariant DNN.$\}$.

As preparation, we define a *stabilizer subgroup* associated with G. In this section, for brevity, we consider that the action G is transitive, i.e. for any $i \in \{1, 2, ..., n\}$, there exists $g \in G$ that satisfies $g \cdot 1 = i$. We define the stabilizer subgroup $\mathrm{St}(G) \subset G$ as $\mathrm{St}(G) = \{g \in G \mid g \cdot 1 = 1\}$. Here, $\mathrm{St}(G)$ is a subgroup of G which fixes the first coordinate. We utilize this subgroup for decomposing equivariant functions and obtain the following bound:

Theorem 3 (Generalization of Equivariant DNN)**.** *Suppose* G *is transitive, and any* $\widetilde{f}^G \in \widetilde{\mathcal{F}}^G = \widetilde{\mathcal{F}}^G(I)$ *is uniformly bounded by* 1*. Then, for any* $\varepsilon > 0$*, there exists constant* $\widetilde{C} > 0$ *that are independent of* n, m *and* ε*, the following inequality holds with probability at least* $1 - 2\varepsilon$*:*

$$\mathcal{G}(\widetilde{\mathcal{F}}^G) \leq \underbrace{\sqrt{\frac{\widetilde{C}}{|\mathrm{St}(G)|\, m^{2/n}}}}_{=:I_1'} + \underbrace{\sqrt{\frac{2\log(2/\varepsilon)}{m}}}_{=:I_2'}.$$

The result shows that equivariant DNNs also achieves the improved generalization bound by the volume of its QFS of $\mathrm{St}(G)$, i.e., the main term satisfies $I_1' \propto \sqrt{\mathcal{N}_{\varepsilon,\infty}(\Delta_{\mathrm{St}(G)})}$. The remainder term I_2' has a smaller order than I_1'. Thus, it is considered to be negligible in our analysis.

By Theorem 3, we obtain the following specific generalization bound with $G = S_n$.

Corollary 2 (Generalization of S_n-equivariant DNN)**.** *Consider the same setting as Theorem 2. Then, for any $\varepsilon > 0$ and sufficiently large n, the following inequality holds with probability at least $1 - 2\varepsilon$:*

$$\mathcal{G}(\widetilde{\mathcal{F}}^{S_n}) \leq \sqrt{\frac{\widetilde{C}}{(n-1)!\, m^{2/n}}} + \sqrt{\frac{2\log(1/2\varepsilon)}{m}}.$$

This corollary describes that S_n-equivariant DNNs can be improved bound by $\sqrt{(n-1)!}$. In Section B in the supplementary material, we relax the transitive setting for G and provide more general results with non-transitive G.

Even with this result, the slow decay rate in m is resolved by the improvement of the bound by n due to invariance. The detailed discussion is similar to that of Remark 2.

Proof sketch for Theorem 3: As a preparation, we define a set of G-equivariant functions with multivariate outputs and their covering numbers. To this end, we reform the G-equivariant function $\widetilde{f}^G : I \to \mathbb{R}^n$ to a combination of $\mathrm{St}(G)$-invariant functions. Proposition 3.1 in Sannai et al. [2019] shows the following formulation:

$$\widetilde{f}^G = (f^{\mathrm{St}(G)} \circ \tau_{1,1}, \cdots, f^{\mathrm{St}(G)} \circ \tau_{1,n})^\top, \qquad (4)$$

where $f^{\mathrm{St}(G)} : I \to \mathbb{R}$ is an existing $\mathrm{St}(G)$-invariant function, and $\tau_{1,j} \in G$ is a linear map for $j = 1, .., n$ such that it makes the first coordinate of an input move to the j-th coordinate. The detailed results are provided in Proposition 10 in Section B. By the representation, we can evaluate a covering number of $\widetilde{\mathcal{F}}^G(I)$ by that of $\mathrm{St}(G)$-invariant functions. For a multi-output function $f : I \to \mathbb{R}^n$ as $f = (f_1, ..., f_n)$, we define a norm $|||f|||_{L^\infty(I)} := \max_{j=1,...,n} \|f_j\|_{L^\infty(I)}$. Also, let $\widetilde{\mathcal{N}}_{\varepsilon,\infty}(\Omega)$ be a covering number of Ω in terms of $|||\cdot|||_{L^\infty(I)}$.

The remaining steps of this proof are similar to those of Theorem 2. □

6 EXPERIMENTAL RESULT

We experimentally validate Theorem 2 by measuring a generalization gap with synthetic data. We consider a regression task to find a sum of n scalars, which is a problem solvable by invariant functions.

We generate synthetic data by the following process. For inputs, we generate $N = nd$ random variables $x_1, ..., x_N$ that are independently and identically generated from a standard normal distribution. We generate an output variable $y = \sum_{i=1}^N x_i$. We regard this as an S_n-invariant function in the following way. We regard $x_1, ..., x_N$ as n d-dimensional vectors $v_1, ..., v_n$ and then we can give the permutation action of S_n on $(v_1, ..., v_n)$. This induces the action of S_n on $(x_1, ..., x_N)$. A function $(x_1, ..., x_N) \mapsto y = \sum_{i=1}^N x_i$ is invariant to the permutation actions of S_n.

We solve the regression problem by *DeepSets* [Zaheer et al., 2017], which is an S_n-invariant DNN with given n. DeepSets consists of S_n-equivariant layers (the first three layers), an S_n-invariant layer, and a fully connected layer (the last layer). A number of units of each layer is as follows: $N \to 128 \to 64 \to 32 \to 32 \to 1$.

We vary $n = 2, 4, 6, 8$ and set $N = 48$, then we consider configurations $(n, d) \in \{(2, 24), (4, 12), (6, 8), (8, 6)\}$ such as satisfying $N = nd$. We generate $m = 60$ samples for training and 10000 samples for testing. We train *DeepSets* with 500 epochs, batch size 4, learning rate 0.001, and the Adam optimizer.

Figure 4 illustrates the result, which shows the mean over five trials with different random seeds.[1] The horizontal axis shows n and the vertical axis shows a logarithm of the generalization gaps. From the result, we can confirm two things. First, the theoretical bound is a certain upper bound of the experimental value by DeepSets. Second, the slope of the experimental value is same to the theoretical slope. This supports our claim that the degree of invariance n reduces the generalization gap.

7 DISCUSSION AND COMPARISON

7.1 EFFECT OF INVARIANCE / EQUIVARIANCE

We identify the effect of invariance / equivariance on the generalization bounds of DNNs. With invariant / equivariant DNNs, the volume of the corresponding QFS decreases, hence the generalization bounds also decrease. In order for the bounds to reflect the volume of the QFS, its convergence rate in m must be slow. However, because the volume of the QFS decays sufficiently fast in n, the generalization bound decays rapidly as n increases. Figure 5 shows a logarithmic version of the bound against m in Figure 1.

7.2 RELATION WITH SOKOLIC ET AL. [2017]

The closest study to our work is the generalization error analysis of invariant classifiers by Sokolic et al. [2017]. The

[1] A reviewer suggested that the experimental result should be expressed in a table. However, since the table is not suitable to express the slope of the gaps with respect to n, we continue to use the figure.

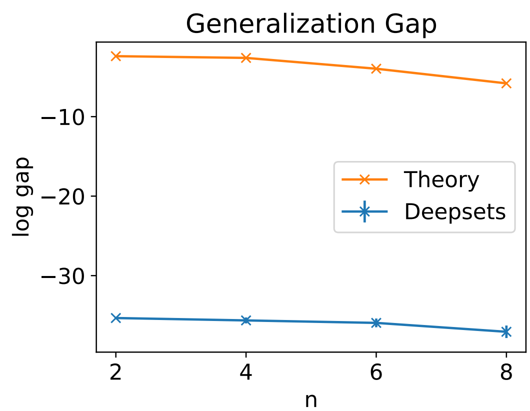

Figure 4: Generalization Gap. Theory bound v.s. experimental result. *Theory* (orange line) denotes $(1/(\sqrt{n!m^{2/n}}))$, and *Deepsets* (blue line) denote the generalization gaps in the experiments with $n \in \{2,4,6,8\}$.

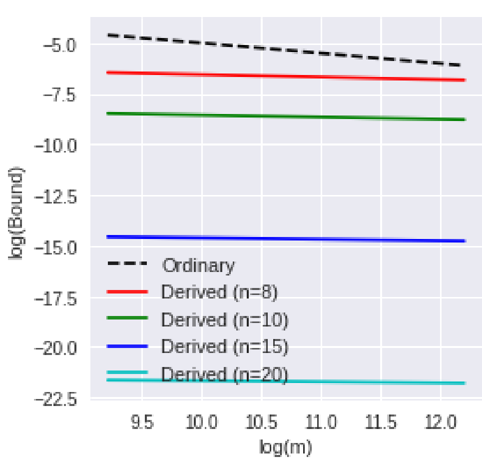

Figure 5: Logarithmic order of the bound for the generalization gap against $\log(m)$. *Ordinary* (dashed line) denotes $\log(1/\sqrt{m})$ without invariance, and *Derived* (colored lines) denotes the bound $\log(1/\sqrt{n!m^{2/n}})$ with $n \in \{8, 10, 15, 20\}$. The bound decreases sharply by n with m increasing to $200,000 \approx \log(12)$.

study shows that a number of *transformations* by invariance describes their generalization bound. While the result is similar to our study, our result improves their analysis in the following two ways.

(I). Our results are valid without the division assumptions in Sokolic et al. [2017]. The study assumes that the input space $\mathcal{X}$ can be written as $\mathcal{X}_0 \times T$ using a set of transformations T and a base space $\mathcal{X}_0$. This assumption is hard to confirm for two reasons. First, the set of transformations is not the group itself in general. For example, consider the trivial action of a symmetric group, then despite the fact that the group is a symmetric group, the set of transformations is a single point set consisting of identities. Thus, we need to calculate the set of transformations on a case-by-case analysis. Second, it is difficult to find the base space. In the graph neural network case [Maron et al., 2018], the action is the permutation of nodes on adjacency matrices. In this case, it is hard to find the base space of this action. In contrast, our result is valid in this case.

(II). Our results relax the stability assumption and achieve an accurate result. Sokolic et al. [2017] places an algorithmic stability assumption that allows them to ignore the complexity of hypothesis spaces and build a theory solely on the complexity of the input space. However, this stability assumption is not satisfied by deep learning in particular [Hoffer et al., 2017, Nagarajan and Kolter, 2019]. In this study, we construct a theory that is independent of stability assumptions by connecting the functional hypothesis space and the theory of QFSs.

7.3 ANALYSIS FOR EXPRESSIVE POWER OF INVARIANT DNNS

We discuss an expressive power of invariant neural networks, which determines how small the empirical loss $R_m(f^G)$ would be. A volume of $R_m(f^G)$ is not the main concern of this study, but it is an important factor for the actual performance.

We investigate the expressive power of invariant deep neural networks. To the aim, we define the *Hölder space*, which is a class of smooth functions.

Definition 4 (Hölder space)**.** Let $\alpha > 0$ be a degree of smoothness. For $f : I \to \mathbb{R}$, the *Hölder norm* is

$$\|f\|_{\mathcal{H}^\alpha} := \max_{\beta:|\beta|<\lfloor\alpha\rfloor} \|\partial^\beta f(x)\|_{L^\infty(I)} + \max_{\beta=\lfloor\alpha\rfloor} \sup_{x,x'\in I, x\neq x'} \frac{|\partial^\beta f(x) - \partial^\beta f(x')|}{\|x-x'\|_\infty^{\alpha-\lfloor\alpha\rfloor}},$$

A B-radius closed ball in the *Hölder space* on I is defined as $\mathcal{H}^\alpha_B = \{f \in \mathcal{H}^\alpha \mid \|f\|_{\mathcal{H}^\alpha} \leq B\}$.

$\mathcal{H}^\alpha$ is a set of bounded functions that are α-times differentiable. The notion is often utilized in characterizing the expressive power of DNNs (e.g., refer to Schmidt-Hieber [2017]). Then, we achieve the following result of an expressive power of invariant DNNs:

Theorem 4 (Approximation rate of invariant DNNs)**.** *For any $\varepsilon > 0$, suppose $\mathcal{F}^{S_n}$ has at most $\mathcal{O}(\log(1/\varepsilon))$ layers and $\mathcal{O}(\varepsilon^{-D/\alpha}\log(1/\varepsilon))$ non-zero parameters. Then, for any S_n-invariant $f^* \in \mathcal{H}_B^\alpha$, there exists $f^{S_n} \in \mathcal{F}^{S_n}$ such that*

$$\|f^{S_n} - f^*\|_{L^\infty(I)} \leq \varepsilon.$$

Theorem 4 clarifies the expressive power of DNNs by showing the sufficient number of parameters to make the error arbitrarily small. This result shows that the error decreases as the number of parameters increase with the rate $-D/\alpha$ up to logarithm factors. Importantly, this rate is the optimal rate *without* invariance [Yarotsky, 2017]. Hence, we prove that invariant DNNs can achieve the optimal approximation rate even with invariance.

8 CONCLUSION

We proposed the generalization theory, which describes the errors and the effect of invariance / equivariance in a quantitative way. We proved that the order of invariance improves the generalization bound. Moreover, we prove S_n-invariant DNNs maintain the high expressive power regardless of the invariance property.

Author Contributions

AS conceived the idea of quotient feature spaces and wrote the paper. MI computed the covering of the function space and wrote the paper. KM created the code.

Acknowledgements

The authors would like to thank the anonymous reviewers for their suggestions and helpful comments. The first author would like to thank Professor Kazuaki Miyatani for sharing his knowledge of pseudo distance induced on quotient spaces. AS was supported t by the JSPS Grant-in-Aid for Scientific Research C (20K03743). MI was supported by JSPS KAKENHI (18K18114) and JST Presto (JPMJPR1852). KM was supported by the JSPS Grant-in-Aid for Early-Career Scientists (20K19930).

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

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

Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In *Advances in Neural Information Processing Systems*, pages 9142–9153, 2019.

Felix A Faber, Alexander Lindmaa, O Anatole Von Lilienfeld, and Rickard Armiento. Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals. *Physical review letters*, 117(13):135502, 2016.

Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

Jan Eric Lenssen, Matthias Fey, and Pascal Libuschewski. Group equivariant capsule networks. In *Advances in Neural Information Processing Systems*, pages 8844–8853, 2018.

Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018a.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, pages 820–830, 2018b.

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

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. *arXiv preprint arXiv:1901.09342*, 2019.

Haggai Maron, Or Litany, Gal Chechik, and Ethan Fetaya. On learning sets of symmetric elements. *arXiv preprint arXiv:2002.08599*, 2020.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5115–5124, 2017.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 11615–11626, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

Michelle Ntampaka, Hy Trac, Dougal J Sutherland, Sebastian Fromenteau, Barnabás Póczos, and Jeff Schneider. Dynamical mass measurements of contaminated galaxy clusters using machine learning. *The Astrophysical Journal*, 831(2):135, 2016.

Siamak Ravanbakhsh. Universal equivariant multilayer perceptrons. *arXiv preprint arXiv:2002.02912*, 2020.

Siamak Ravanbakhsh, Junier B Oliva, Sebastian Fromenteau, Layne Price, Shirley Ho, Jeff G Schneider, and Barnabás Póczos. Estimating cosmological parameters from the dark matter distribution. In *ICML*, pages 2407–2416, 2016.

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

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

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

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

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

John Shawetaylor. Sample sizes for threshold networks with equivalences. *Information and Computation*, 118(1): 65–72, 1995.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel Rodrigues. Generalization error of invariant classifiers. In *Artificial Intelligence and Statistics*, pages 1094–1103. PMLR, 2017.

Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2530–2539, 2018.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.

Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

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

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in neural information processing systems*, pages 4800–4810, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.