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

Akiyoshi Sannai\textsuperscript{1,2} Masaaki Imaizumi\textsuperscript{3,1} Makoto Kawano\textsuperscript{3}

\textsuperscript{1}RIKEN Center for Advanced Intelligence Project, Chuo, Tokyo, Japan
\textsuperscript{2}Keio University, Minato, Tokyo, Japan
\textsuperscript{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, Niampaka 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, Lentssen 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, Niampaka 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 generalization 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
We have mainly two technical contributions. First, we define $$G$$ which is regarded as a feature space associated with 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 $$\phi_G: \mathbb{R}^n \rightarrow \mathbb{R}^p / G$$ 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 $$\Delta_G$$ (invariant case) or $$\Delta_{St(G)}$$ (equivariant case), where $$St(G) \subset G$$ 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:

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

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 $$\varepsilon > 0$$,

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

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 outweights 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.

### 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.
We provide a general concept of the invariance and equivariance on a finite group \( G \), which is isomorphic as metric spaces. The supplementary material maintains all full proofs.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Deep Network} & \text{Group} & \text{vol}(\Delta_G) & \text{vol}(\Delta_{G^e}(G)) \\
\hline
\text{Deep Sets} & S_n & O(1/(n!)) & O(1/(1!)) \\
\text{G-CNN} & C_4 & O(1/4) & O(1) \\
\text{Graph Network} & S_n & O(1/((\# \text{of nodes}) - 1)!)) & O(1/((\# \text{of nodes}) - 1)!)) \\
\text{Tensor Network} & G & O(1/G) & O(1/\text{St}(G)) \\
\text{DSS} & S_n \times G'(G' \subset S_N) & O(1/(n!|G'|)) & O(1/(n-1)!|\text{St}(G')|) \\
\hline
\end{array}
\]

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 \( \text{vol}(\Delta_G) = \mathcal{C}_{\text{vol}}(\Delta_G) \), where \( \mathcal{C}_{\text{vol}}(\Delta_G) \) is a covering number of \( \Delta_G \) 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 \( f : \Omega \rightarrow \mathbb{R} \) with a set \( \Omega \), \( \| f \|_{L^q} := \left( \int_{\Omega} |f(x)|^q \, dx \right)^{1/q} \) denotes the \( L^q \)-norm for \( q \in [0, \infty] \). For a subset \( \Lambda \subset \Omega \), \( f_{|\Lambda} \) denotes a restriction of \( f \) to \( \Lambda \). \( C(\Omega) \) denotes a set of continuous functions on \( \Omega \). For an integer \( z \), \( z! = \prod_{j=1}^{z} j \) denotes a factorial of \( z \). For a set \( \Omega \), \( \text{id}_\Omega \) denotes the identity map on \( \Omega \), namely \( \text{id}_\Omega(x) = x \) for any \( x \in \Omega \). For a subset \( \Delta \subset \mathbb{R}^n \), \( \text{int}(\Delta) \) denotes a set of inner points of a set \( \Delta \). For metric spaces \( \Delta \) and \( \Delta' \), \( \Delta \cong \Delta' \) 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 \( f : \mathbb{R}^n \rightarrow \mathbb{R}^M \) is

- **\( G \text{-invariant}** if \( f(g \cdot x) = f(x) \) holds for any \( g \in G \) and any \( x \in \mathbb{R}^n \),
- **\( G \text{-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 \( \Omega \), \( C^G(\Omega) \) denotes a set of \( G \text{-invariant} \) an continuous functions on \( \Omega \).

We formulate deep neural networks (DNNs) with invariance and equivariance. In this study, we consider fully connected DNNs with the ReLU activation function \( \text{ReLU}(x) = \max(0, x) \). Let us consider a layer-wise map \( Z_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_{i+1}+1} \) defined by \( Z_i(x) = \text{ReLU}(W_i x + b_i) \), where \( W_i \in \mathbb{R}^{d_{i+1} \times d_i} \) and \( b_i \in \mathbb{R}^{d_{i+1}} \) for \( i = 1, \ldots, H \). 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

\[
 f(x) := Z_H \circ Z_{H-1} \circ \ldots \circ Z_2 \circ Z_1(x). 
\]

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 \text{-invariant} / \text{equivariant} DNN**, if \( f \) is a **\( G \text{-invariant} / \text{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 $Z_1, \ldots, Z_j$, a permutation-equivariant linear layer $Z_L$, and a fully-connected layer $Z_F$. Each equivariant layer maintains a parameter matrix $W_i = \lambda I + \gamma I^1$, $\lambda, \gamma \in \mathbb{R}, I = [1, \ldots, 1]^T$, which makes $Z_j$ be equivariant. Then, a DNN $f = Z_F \circ Z_L \circ Z_1 \circ \cdots Z_j$ is a permutation-equivariant 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 a higher-order tensors. The study considered a tensor $W \in \mathbb{R}^{d \times \ldots \times d}$ and an action $g \in G$ on the tensor as $(g \cdot W)_{i_1 \ldots i_k} = W_{g^{-1}(i_1) \ldots g^{-1}(i_k), j}$, for $i_k = 1, \ldots, n, k' = 1, \ldots, k$, and $j = 1, \ldots, a$. 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 $\sup_{y' \in G} L(y(y')) \leq 1$ and 1-Lipschitz continuous. Let $P^* (x, y)$ be a distribution on $\mathbb{R}^M$ 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 $\{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ 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 $\mathbb{E}_f (R(f))$ 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

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

named a quotient map. By the definition of $g$, such $\phi_G$ 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 $\Delta_G$, 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 $\Delta$ using a covering number $\mathcal{N}_{\mathbb{E}, \omega}(\Delta) := \inf \{ N \mid \exists \{ x_j \}_{j=1}^N \subset \Delta, \text{ s.t. } \cup_{j=1}^N \{ x \mid \| x - x_j \|_2 \leq \epsilon \} \supset \Delta \}.$

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 $\phi_{S_n}$ and $\Delta_{S_n}$. With the case, an action $\sigma \in S_n$ is a permutation of indexes of $x = (x_1, \ldots, x_n) \in I$. For $i = 1, \ldots, n$, we define a map $\max_i (\{ x_1, \ldots, x_n \})$ which returns the $i$-th largest element of $\{ x_1, \ldots, x_n \}$.

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 $\phi(x) := (\max_1 (\{ x_1, \ldots, x_n \}), \ldots, \max_n (\{ x_1, \ldots, x_n \})).$ Then, we obtain $\phi (\Delta) = \Delta_{S_n}$. Figure 2 illustrates $\Delta_{S_n}$ for some $n$. Intuitively, any element of $I$ corresponds to some element of $\Delta_{S_n}$ with an existing action $\sigma \in S_n$, namely, $I = \cup_{\sigma \in S_n} \{ \sigma \cdot x \mid x \in \Delta \}.$ With the help of the explicit formulation of $\Delta_{S_n}$, 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 $\Delta_{S_n}$). There is a constant $C$ such that for small enough $\epsilon > 0$, we obtain

$$\mathcal{N}_{\mathbb{E}, \omega}(\Delta_{S_n}) \leq C / (n! \cdot \epsilon^n).$$

Lemma 2 provides an important claim: the volume of $\Delta_{S_n}$ is proportional to $1 / (n!)$, i.e., the volume significantly decreases with the increases in $n$. The term $\epsilon^{-n}$ is usual for covering numbers, i.e., $\mathcal{N}_{\mathbb{E}, \omega}(I) \leq C / \epsilon^n$ 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 $\Delta_G$ 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 $\{ g_1, \ldots, g_k \mid g_k \in G \}$ is defined as a complete
system of representatives of $H \setminus G$ if $K = |G|/|H|$ and $G = \bigcup_{k=1}^{K} H \cdot g_k$ 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 $\{g_1, \ldots, g_K\} \subset S_n, k = 1, \ldots, K$ be a complete system of representatives of $G \setminus S_n$. Then, $\Delta_k \cong \Delta_{S_n}$ holds as metric spaces for all $k = 1, \ldots, K$. Furthermore, its induced set $\Delta_G := \bigcup_{k=1}^{K}|G| \cdot \Delta_k$ satisfies $\phi_G(\Delta_G) = \Delta_G$.

Proposition 3 shows that we can describe $\Delta_G$ by $\tilde{\Delta}_G$, which is a combination of complete systems of representatives of $G \setminus S_n$. Intuitively, we can define $\Delta_G$ by a union of several transformed $\Delta_{S_n}$.

We describe an example with $n = 3$ and $G = S_2$. A complete system of representatives of $S_2 \setminus S_3$ 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 $\Delta_{S_3} = \Delta_{S_2}$. Then, we can represent $\Delta_{S_2}$ by $\Delta_{S_3}$ 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, $\Delta_{S_2}$ is a union of $\Delta_{S_3}$ ($= g_1 \cdot \Delta_{S_3}$), reflected $\Delta_{S_3}$ ($= g_2 \cdot \Delta_{S_3}$), and rotated $\Delta_{S_3}$ ($= g_3 \cdot \Delta_{S_3}$).

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

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

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

Similar to Lemma 1, the result of Lemma 2 states that the covering volume of $\Delta_G$ 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 $\Delta_G$. Namely, we show that a covering number of a set $\mathcal{F}^G(I) = \{f : I \to \mathbb{R}^M \mid f$ is a $G$-invariant DNN$\}$ is evaluated by comparison with the volume of $\Delta_G$. We also define $\mathcal{F}(\Delta_G) := \{f : \Delta_G \to \mathbb{R}^M \mid f$ 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.** $\phi_G$ induces a bijection $\hat{\phi}_G : C(\Delta_G) \to C^G(I)$.

Further, $f \in C(\Delta_G)$ is $K$-Lipschitz continuous if and only if $\phi_G(f)$ 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 $\Delta_G$ without invariance.

Finally, we evaluate the volume of $\mathcal{F}(\Delta_G)$ in terms of the volume of $\Delta_G$. We provide an inequality which bounds the covering number of $\mathcal{F}(\Delta_G)$ by a polynomial of the volume of $\Delta_G$, whereas the commonly used inequality only includes the logarithm of the volume of $\Delta_G$, 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 $C_\Delta$-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

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

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

**Remark 1** (Linear bound in $\mathcal{N}_{\delta, \infty}(\Delta_G)$). 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 $\mathcal{N}_{\delta, \infty}(\Delta_G)$. In general, log of $\mathcal{N}_{\delta, \infty}(\mathcal{F}(\Delta_G))$ 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 $\Delta_G$, as a value related to invariance and achieve the linear bound in $\mathcal{N}_{\delta, \infty}(\Delta_G)$.

### 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 $\varepsilon$, and the following inequality holds with probability at least $1 - 2\varepsilon$:

$$\mathbb{E}_I \left[ \frac{C}{|G|^{m^2/n}} + \frac{2 \log(1/2\varepsilon)}{m} \right].$$

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 $1 - 2\varepsilon$:

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

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 $1 - 2\varepsilon$:

$$G(\mathcal{F}^G) \leq \sqrt{\frac{2\log(1/2\varepsilon)}{m}}$$

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

Second, we bound the term $\log \mathcal{N}_{G,\infty}(\mathcal{F}^G(T))$ in (1) by $\log \mathcal{N}_{G,\infty}(\mathcal{F}(G))$ by using the result in Proposition 4. This enables us to evaluate the error with $G$-invariance using $G$.

Third, we bound $\log \mathcal{N}_{G,\infty}(\mathcal{F}(G))$ by the term with $\mathcal{N}_{G,\infty}(\mathcal{F})$. 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 $\tilde{\mathcal{F}}^G(I) = \{ f : I \rightarrow \mathbb{R}^p \mid f$ 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 $i \in \{1, 2, \ldots, n\}$, there exists $g \in G$ that satisfies $g \cdot 1 = i$. We define the stabilizer subgroup $\text{St}(G) \subseteq G$ as $\text{St}(G) = \{ g \in G \mid g \cdot 1 = 1 \}$. Here, $\text{St}(G)$ 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 $\tilde{f}^G \in \tilde{\mathcal{F}}^G = \mathcal{F}^G(I)$ is uniformly bounded by 1. Then, for any $\varepsilon > 0$, there exists constant $\tilde{C} > 0$ that are independent of $n, m, \varepsilon$, the following inequality holds with probability at least $1 - 2\varepsilon$:

$$G(\tilde{\mathcal{F}}^G) \leq \sqrt{\frac{\tilde{C}}{|\text{St}(G)| m^2/n}} + \sqrt{\frac{2\log(1/2\varepsilon)}{m}}.$$

The result shows that equivariant DNNs also achieves the improved generalization bound by the volume of its QFS of $\text{St}(G)$, i.e., the main term satisfies $I_1' \approx \sqrt{\mathcal{N}_{G,\infty}(\text{St}(G))}$. 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 $G = S_n$.

![Figure 3: Illustration of $\Delta_{S_2} = \bar{\Delta}_{S_2} = g_1 \cdot \Delta_{S_3} \cup g_2 \cdot \Delta_{S_3} \cup g_3 \cdot \Delta_{S_3}$. The blue cube is $I$, and the green polyhedrons are $\Delta_{S_2}$ and $g_k \cdot \Delta_{S_3}, k = 1, 2, 3.$](image-url)
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 $1 - 2\varepsilon$:

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

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 $\bar{f}^G : I \to \mathbb{R}^n$ to a combination of $\text{St}(G)$-invariant functions. Proposition 3.1 in Sannai et al. [2019] shows the following formulation:

$$\bar{f}^G = (f^{\text{St}(G)} \circ \tau_{1,1}, \ldots, f^{\text{St}(G)} \circ \tau_{1,n})^\top, \quad (4)$$

where $f^{\text{St}(G)} : I \to \mathbb{R}$ is an existing $\text{St}(G)$-invariant function, and $\tau_{1,j} \in G$ is a linear map for $j = 1, \ldots, n$ 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 $\mathcal{F}^G(I)$ by that of $\text{St}(G)$-invariant functions. For a multi-output function $f : I \to \mathbb{R}^n$ as $f = (f_1, \ldots, f_n)$, we define a norm $\|f\|_{L_{\infty}(I)} := \max_{j = 1, \ldots, n} \|f_j\|_{L_{\infty}(I)}$. Also, let $\mathcal{N}_{\varepsilon,n}(\Omega)$ be a covering number of $\Omega$ 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 $x_1, \ldots, x_N$ 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 $x_1, \ldots, x_N$ as $n \cdot d$-dimensional vectors $v_1, \ldots, v_n$ and then we can give the permutation action of $S_n$ on $(v_1, \ldots, v_n)$. This induces the action of $S_n$ on $(x_1, \ldots, x_N)$. A function $(x_1, \ldots, x_N) \mapsto y = \sum_{i=1}^{N} x_i$ 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. The horizontal axis shows $n$ and the vertical axis shows a logarithm of the generalization gap. 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,

![Generalization Gap](image)

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 $n \in \{2, 4, 6, 8\}$. 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.
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

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

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

$\mathcal{H}^\alpha$ is a set of bounded functions that are $\alpha$-times differentiable. The notion is often utilized in characterizing the expression 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 $O((\log(1/\varepsilon))$ layers and $O(e^{-D/\alpha} \log(1/\varepsilon))$ 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 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 Cordnier. 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.

Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. arXiv preprint arXiv:1805.07091, 2018.
A SUPPLEMENT

We provide the deferred proofs of each section.

A.1 PROOF FOR SECTION 3

Proof of Lemma[4] For any $\varepsilon > 0$, there is a sequence of rational numbers $\{p_i/q_i\}$ such that $p_i/q_i < \varepsilon$ and converges to $\varepsilon$. Assume that Lemma[1] holds for rational numbers, then we have $\mathcal{N}_\varepsilon(\Delta_n) \leq \mathcal{N}_\varepsilon(p_i/q_i,\Delta_n) \leq C/(n! (p_i/q_i)^n)$. Since $1/x^i$ is a continuous function and $\{p_i/q_i\}$ converges to $\varepsilon$, we obtain $\mathcal{N}_\varepsilon(\Delta_n) \leq C/((n! \varepsilon)^n)$. Hence it is enough to show the case of rational numbers.

We assume $\varepsilon = p/q$ for some integers $p, q > 0$. Let $\mathcal{C}(I)$ be the covering of $I$, which is a set of $\varepsilon$-cubes

$$c_{j_1...j_n} = \{x = (x_i) \in I \mid x_i \in [x_i \varepsilon(j_i) \leq x_i \varepsilon(j_i + 1),
$$

for $j_i = 1,...,[q/p] + 1$. We can easily see that $\mathcal{C}(I)$ attains the minimum number of $\varepsilon$-cubes covering $I$ and the number is $(\varepsilon^{-1} + 1)^n = \frac{\varepsilon^n}{\pi^2} + \mathcal{O}(\varepsilon^{-(n-1)})$. We show that we can find a subset of $\mathcal{C}(I)$ which cover $\Delta_n$, and whose cardinality is $\frac{\varepsilon^n}{\pi^2}$ + $\mathcal{O}(\varepsilon^{-(n-1)})$. The proof is as follows. At first, we calculate the number $A$ of cubes in $\mathcal{C}(I)$ which intersect with the boundary of $\sigma \cdot \Delta$. Then since the number of the orbit of the cubes which do not intersect with the boundary of $\sigma \cdot \Delta$ is $n!$, if $A$ is $\mathcal{O}(\varepsilon^{-(n-1)})$, we can find the covering whose cardinality is $\frac{\varepsilon^n}{\pi^2} + \mathcal{O}(\varepsilon^{-(n-1)})$. Since $\sigma \cdot \Delta$ is $\{x \in I \mid x_{\sigma^{-1}(1)} \geq x_{\sigma^{-1}(2)} \geq \cdots \geq x_{\sigma^{-1}(n)}\}$, any boundary of $\sigma \cdot \Delta$ is of the form $\{x \in I \mid x_{\sigma^{-1}(1)} = \cdots = x_{\sigma^{-1}(i)} = x_{\sigma^{-1}(i+1)} = \cdots = x_{\sigma^{-1}(n)}\}$.

From here, we fix $\sigma$ and $i$. Consider the canonical projection $\pi: \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$ which sends $x_{\sigma^{-1}(i)}$-axis to zero. $\pi$ induces the map $\widetilde{\pi}: \mathcal{C}(I) \rightarrow \mathcal{C}(\pi(I))$, where $\mathcal{C}(\pi(I))$ is the covering of $\pi(I)$. Let $\mathcal{C}(I)_B$ denote the subset of cubes in $\mathcal{C}(I)$ which intersect with the set $B = \{x \in I \mid x_{\sigma^{-1}(i)} = x_{\sigma^{-1}(i+1)}\}$. Then we can see that $\widetilde{\pi}$ is injective on $\mathcal{C}(I)_B$. Assume that there are two cubes in $\mathcal{C}(I)_B$ whose images by $\widetilde{\pi}$ are equal. Let us denote the centers of two cubes by $c_{j_1...j_n}$ and $c_{k_1...k_n}$. Then, $\sigma$ only kills $x_{\sigma^{-1}(i)}$, $j_\sigma = k_\sigma$ holds for $p \neq \sigma^{-1}(i)$. But since $c_{j_1...j_n}$ and $c_{k_1...k_n}$ are in $\mathcal{C}(I)_B$, we have $j_{\sigma^{-1}(i)} = k_{\sigma^{-1}(i)}$ and $k_{\sigma^{-1}(i)} = k_{\sigma^{-1}(i+1)}$. Hence $j_\sigma = k_\sigma$ for any $p$ and $\widetilde{\pi}$ is injective on $\mathcal{C}(I)_B$.

Next, let $\mathcal{C}(I)_\tilde{B}$ be the subset of $\varepsilon$-cubes in $\mathcal{C}(I)$ which intersect a boundary of $\sigma \cdot \Delta$ for some $\sigma$. We see that the cardinality of $\mathcal{C}(I)_\tilde{B}$ is bounded by $c \varepsilon^{-(n-1)}$ for some constant $c > 0$. Since the number of components of the boundaries is finite, we prove the claim for a component $B$. As we see before, $\widetilde{\pi}\mid\mathcal{C}(I)_\tilde{B}$ is injective. This result implies that the number of cubes that intersect $B$ is bounded by a number of $\varepsilon$-cubes in $\mathcal{C}(p(I)) = \mathcal{C}(I) \setminus \mathcal{C}(I)_\tilde{B}$. Then we note that the action of $\mathcal{S}_n$ on $\mathcal{C}(I)_F$ is free, namely the number of the orbit of any cube in $\mathcal{C}(I)_F$ is $|\mathcal{S}_n|$. Hence,

$$|\mathcal{C}(I)_F \cap \Delta| = \{c \in \mathcal{C}(I) \mid c \subset \Delta\} = 1/|\mathcal{S}_n||\mathcal{C}(I)_F| \leq \varepsilon^{-n}/|\mathcal{S}_n|.$$

Here, $(\mathcal{C}(I)_F \cap \Delta) \cup \mathcal{C}(I)_\tilde{B}$ gives the covering of $\Delta$. This covering gives

$$\mathcal{N}_{\mathcal{S}_n}(\Delta) \leq \frac{\varepsilon^{-n}}{n!} + c \varepsilon^{-(n-1)}.$$


Proof of Proposition 2. We confirm that $\hat{\phi}_G$ satisfies both the conditions 1 and 2. As the action of $G$ preserves the distance, $g_k : \Delta_{S_n} \to g_k \cdot \Delta_{S_n}$ is an isomorphism on metric spaces. Hence, condition 1 is satisfied.

For condition 2, we consider $y \in \Delta_G = \phi_G(I)$. Then, there is an element $x \in I$ such that $y = \phi_G(x)$. As $I = \bigcup_{\sigma \in S_n} \sigma \cdot \Delta_{S_n}$, there exist $\sigma \in S_n$ and $z \in \Delta_{S_n}$ such that $x = \sigma \cdot z$.

In contrast, as $\{g_1, \ldots, g_k | g_k \in G\}$ is a complete system of representatives of $G \setminus S_n$, there exist $\tau \in G$ and $g_k$ such as $\sigma = \tau \cdot g_k$. Then $\phi_G(g_k z) = \phi_G(\tau \cdot g_k z) = \phi_G(x) = y$ and $g_k \cdot z \in \Delta_G$. Hence $\phi_G(\Delta_G) = \Delta_G$.

A.2 Proof for Section 3.2

Proof of Proposition 4. We prove $\hat{\phi}_G$ is injective and surjective. Assume $f \in C(\Delta_G)$ and put $\hat{\phi}_G(f) = f \circ \phi_G$. Then since $\phi_G$ is $G$-invariant, so is $\hat{\phi}_G(f)$. Also, since $\phi_G$ is surjective, $\hat{\phi}_G$ is injective. Take $g \in C^G(I)$, then we define $f \in C(\Delta_G)$ as follows: for any $y \in \Delta_G$, take $x \in I$ such that $\phi_G(x) = y$ and define $f(y) = g(x)$. This map is well defined because $g$ is $G$-invariant. $\hat{\phi}_G(f)(x) = f \circ \phi_G(x) = f(y) = g(x)$. Hence, we obtain the desired result.

Next, we prove the Lipschitz properties. Take $f \in C(\Delta_G)$ and assume $f$ is $K$-Lipschitz. Then for any $x, x' \in I$, $d_G(\phi_G(x), \phi_G(x')) \geq Kd(f(\phi_G(x)), f(\phi_G(x')))$, by $K$-Lipschitz property of $f$. By the definition of $d_G$, we have $d_G(\phi_G(x), \phi_G(x')) \leq d(x, x')$. Hence, $\hat{\phi}_G(f)$ is $K$-Lipschitz continuous. Conversely, assume $\hat{\phi}_G(f)$ is $K$-Lipschitz. Take any $y, y' \in I$, then for any $x, x' \in I$ satisfying $\phi_G(x) = y, \phi_G(x) = y'$, $d(x, x') \geq Kd(f(\phi_G(x)), f(\phi_G(x'))) = d(f(y), f(y'))$, by $K$-Lipschitz property of $\hat{\phi}_G(f)$. Hence by taking infimum of the left hand side, we have $d_G(y, y') = \inf d_G(\phi_G(x), \phi_G(x')) \geq Kd(f(y), f(y'))$. Hence, $f$ is $K$-Lipschitz.

Proof of Proposition 2. We first note that $\phi$ is the identity map on $\Delta$, because elements in $\Delta$ are sorted. This implies $\Delta \cong \phi(\Delta)$. Therefore, it is sufficient to show $\Delta \cong \Delta_{S_n}$. As $\Delta$ is a subset of $I$, we have the distance preserving map $\phi_{S_n, I} : \Delta \to \Delta_{S_n}$.

Then, we show that $\phi_{S_n, I}$ is a bijection. Injectivity: Let us take any $x, y \in \Delta$ such that $\phi_{S_n, I}(x) = \phi_{S_n, I}(y)$. Then $x = g \cdot y$ for some $g \in S_n$. However, as $y$ is in $\Delta$, $\{g \cdot y | g \in S_n\} \cap \Delta = \{y\}$. Hence, $x = y$. Surjectivity: For any $z \in \Delta_{S_n}$, then there is $x \in I$ such that $z = \phi_G(x)$. By the construction of $\Delta$, there is $g \in S_n$ and $y \in \Delta$ that satisfies $x = g \cdot y$. Hence, $z = \phi_{S_n}(x) = \phi_{S_n}(g \cdot y) = \phi_{S_n}(y)$.

Proof of Proposition 5. Firstly, we show $\hat{\phi}_G^{-1}(f) \in \mathcal{F}(I)$ with any $f \in \mathcal{F}^G(I)$. For $f \in \mathcal{F}^G(I)$, we consider $\hat{\phi}_G^{-1}(f) \in C(\Delta_G)$ as Proposition 4. Suppose $f$ and $f'$ are $K$-Lipschitz continuous, then $\hat{\phi}_G^{-1}(f)$ is also $K$-Lipschitz continuous by Proposition 4. Since Zhang et al. [2018] states that Lipschitz continuous functions are represented by DNNs, we have $\hat{\phi}_G^{-1}(f) \in \mathcal{F}(G)$. Fix $f_1, f_2 \in \mathcal{F}^G(I)$. Then, there exist $f'_1, f'_2 \in \mathcal{F}(\Delta_G)$ such that $f_1 = \hat{\phi}_G(f'_1)$ and $f_2 = \hat{\phi}_G(f'_2)$. Then, we have $\|f_1 - f_2\|_{L^\infty(I)} = \|\hat{\phi}_G(f'_1) - \hat{\phi}_G(f'_2)\|_{L^\infty(I)} = \|f'_1 \circ \phi_G - f'_2 \circ \phi_G\|_{L^\infty(I)} \leq \|f'_1 - f'_2\|_{L^\infty(\Delta_G)}$.

Based on the result, we can bound $\mathcal{N}_{\epsilon, \infty}(\mathcal{F}^G(I))$ by $\mathcal{N}_{\epsilon, \infty}(\mathcal{F}(\Delta_G))$. Let us define $N := \mathcal{N}_{\epsilon, \infty}(\mathcal{F}(\Delta_G))$. Then, there exist $f'_1, \ldots, f'_N$ such that for any $f' \in \mathcal{F}(\Delta_G)$, there exists $j \in \{1, \ldots, N\}$ such as $\|f' - f'_j\|_{L^\infty(\Delta_G)} \leq \epsilon$. Here, for any $f \in \mathcal{F}^G(I)$, there exists $f_j := \hat{\phi}_G^{-1}(f'_j) \in \mathcal{F}^G(I)$ and it satisfies $\|f - f_j\|_{L^\infty(I)} \leq \|\hat{\phi}_G(f) - \hat{\phi}_G(f_j)\|_{L^\infty(\Delta_G)} \leq \epsilon$. Then, we obtain the statement.

Proof of Theorem 2. Combining Proposition 5 and 4, we obtain a bound for log $\mathcal{N}_{2\epsilon, \delta, \infty}(\mathcal{F}^G(I))$. Then, we substitute it into 3 and obtain the statement of Theorem 2.
A.3 PROOF FOR SECTION 4

Proof of Proposition 6. We bound a covering number of a set of \( C_A \)-Lipschitz continuous functions on \( \Delta \). Let \( \{x_1, \ldots, x_K\} \subset \Delta \) by a set of centers of \( \delta \)-covering set for \( \Delta \). By Lemma 1, we set \( K = C/(|G| \delta^n) \) with \( \delta \) a parameter \( \delta > 0 \), where \( C > 0 \) is a constant.

We will define a set of vectors to bound the covering number. We define a discretization operator \( A : \mathcal{F}(\Delta_G) \to \mathbb{R}^K \) as

\[
Af = \binom{f(x_1)/\delta, \ldots, f(x_K)/\delta}{\top}.
\]

Let \( \mathcal{B}_\delta(x) \) be a ball with radius \( \delta \) in terms of the \( \| \cdot \|_\infty \)-norm. For two functions \( f, f' \in \mathcal{F}(\Delta_G) \) such as \( Af = A f' \), we obtain

\[
\| f - f' \|_{L^\infty(I)} = \max_{k=1, \ldots, K} \sup_{x \in \mathcal{B}_\delta(x_k)} |f(x) - f'(x)|
\leq \max_{k=1, \ldots, K} \sup_{x \in \mathcal{B}_\delta(x_k)} |f(x) - f(x_k)| + |f'(x_k) - f(x_k)|
\leq 2C_A \delta,
\]

where the second inequality follows \( f(x_k) = f'(x_k) \) for all \( k = 1, \ldots, K \) and the last inequality follows the \( C_A \)-Lipschitz continuity of \( f \) and \( f' \). By the relation, we can claim that \( \mathcal{F}(\Delta_G) \) is covered by \( 2C_A \delta \) balls whose center is characterized by a vector \( b \in \mathbb{R}^K \) such as \( b = Af \) for \( f \in \mathcal{F}(\Delta_G) \). Namely, \( \mathcal{N}_{2C_A \delta, \infty}(\mathcal{F}(\Delta_G)) \) is bounded by a number of possible \( b \).

Then, we construct a specific set of \( b \) to cover \( \mathcal{F}(\Delta_G) \). Without loss of generality, assume that \( x_1, \ldots, x_K \) are ordered satisfies such as \( \|x_k - x_{k+1}\|_\infty \leq 2\delta \) for \( k = 1, \ldots, K - 1 \). By the definition, \( f \in \mathcal{F}(\Delta_G) \) satisfies \( \| f \|_{L^\infty(I)} \leq B, b_1 = f(x_1) \) can take values in \([-B/\delta, B/\delta]\). For \( b_2 = f(x_2) \), since \( \|x_1 - x_2\|_\infty \leq 2\delta \) and hence \( |f(x_1) - f(x_2)| \leq 2C_A \delta \), a possible value for \( b_2 \) is included in \([b_1 - 2\delta/\delta, b_1 + 2\delta/\delta]\). Hence, \( b_2 \) can take a value from an interval with length 4 given \( b_1 \). Recursively, given \( b_k \) for \( k = 1, \ldots, K - 1 \), \( b_{k+1} \) can take a value in an interval with length 4.

Then, we consider a combination of the possible \( b \). Simply, we obtain the number of vectors is \((2cB/\delta) \cdot (4c)^{K-1} \leq (8c^2B/\delta)^{K-1}\) with a universal constant \( c \geq 1 \). Then, we obtain that

\[
\log \mathcal{N}_{2C_A \delta, \infty}(\mathcal{F}(\Delta_G)) \leq (K - 1) \log (8c^2B/\delta).
\]

Then, we specify \( K \) which describe a size of \( \Delta \) through the set of covering centers.

\[\square\]

A.4 PROOF FOR SECTION 5

Proposition 7. Suppose \( G \) is transitive. Then, for any \( \varepsilon > 0 \), we have

\[
\mathcal{N}_{\mathcal{F}(\Delta), \infty}(\mathcal{F}^{G^1}(I)) \leq \mathcal{N}_{\mathcal{F}(\Delta), \infty}(\mathcal{F}^{\text{St}(G)}(I)).
\]

Proof of Proposition 7. The first statement simply follows Proposition 1 with setting \( J = 1 \), since \( g \in G \) is transitive. In the case of \( S_n \), we have \( J = 1 \) and \( \text{Stab}(1) \cong S_{n-1} \). This gives the second statement.

\[\square\]

Proof of Theorem 3 and Corollary 2. For Theorem 3 we combine the bound (5), Lemma 2 and Proposition 5 thus, we obtain the statement.

For Corollary 2 since \( S_n \) is transitive, the statement obviously holds with \(|\text{St}(G)| = |S_{n-1}| = (n - 1)!|\).

\[\square\]

A.5 PROOF FOR SECTION 6

To prove Theorem 4, we consider a Sort map and show that DNNs can represent the map. Let \( \max^{(k)}(x_1, \ldots, x_n) \) be a map which returns the \( k \)-th largest value of inputted elements \( x_1, \ldots, x_n \) for \( k = 1, \ldots, n \). Then, we provide a form of Sort as

\[
\text{Sort}(x_1, \ldots, x_n) = (\max^{(1)}(x_1, \ldots, x_n), \ldots, \max^{(n)}(x_1, \ldots, x_n)).
\]

To represent it, we provide the following propositions.
Proposition 8. \( \max^{(j)}(z_1, \ldots, z_N) \) and \( \min^{(j)}(z_1, \ldots, z_N) \) are represented by an existing deep neural network with an ReLU activation for any \( j = 1, \ldots, N \).

Proof of Proposition 8 First, since
\[
\max(z_1, z_2) = \max(z_1 - z_2, 0) + z_2,
\]
and
\[
\min(z_1, z_2) = -\max(z_1 - z_2, 0) + z_1
\]
hold, we see the case of \( j = 1, N = 2 \). By repeating \( \max(z_1, z_2) \), we construct \( \max^{(1)}(z_1, \ldots, z_N) \) and \( \min^{(1)}(z_1, \ldots, z_N) \). Namely, we prove the claim in the case of \( j = 1 \) and arbitrary \( N \). At first, we assume \( N \) is even without loss of generality, then we divide the set \( \{ z_1, \ldots, z_N \} \) into sets of pairs \( \{ (z_1, z_2), \ldots, (z_{N-1}, z_N) \} \). Then, by taking a max operation for each of the pairs, we have \( \{ y_1 = \max(z_1, z_2), \ldots, y_{N/2} = \max(z_{N-1}, z_N) \} \). We repeat this process to terminate. Then we have \( \max^{(1)}(z_1, \ldots, z_N) \), which is represented by an existing deep neural network. Similarly, we have \( \min^{(1)}(z_1, \ldots, z_N) \). Finally, we prove the claim on \( j = 2, \ldots, N \) by induction. Assume that for any \( N \) and \( \ell < j \), \( \max^{(\ell)}(z_1, \ldots, z_N) \) is represented by a deep neural network. We construct \( \max^{(j)}(z_1, \ldots, z_N) \) as follows: since
\[
\max^{(j-1)}(z_{-\ell}) = \begin{cases} 
\max^{(j-1)}(z_{1}, \ldots, z_{N}) & \text{if } z_\ell \leq \max^{(j)}(z_{1}, \ldots, z_{N}) \\
\max^{(j)}(z_{1}, \ldots, z_{N}) & \text{otherwise} 
\end{cases}
\]
holds, we have \( \max^{(j)}(z_1, \ldots, z_N) = \min(\{ \max^{(j-1)}(z_{\ell}) | \ell = 1, \ldots, N \}) \). By inductive hypothesis, the right hand side is represented by a deep neural network.

Further, we provide the following result for a technical reason.

Proposition 9. The restriction map
\[
\Lambda: \mathcal{F}^S_n(I) \to \mathcal{F}(\Delta_{S_n})
\]
is bijective, where \( \Lambda(f) = f_{\mid \Delta_{S_n}} \).

Proof of Proposition 9 To show the Proposition, we firstly define sorting layers which is an \( S_n \)-invariant network map from \( I \) to \( \Delta \). Then by Proposition 8, \( \text{Sort}(x_1, \ldots, x_n) \) is also a function by an \( S_n \)-invariant deep neural network and \( \text{Sort}(x_1, \ldots, x_n) \) is the function from \( f \) to \( \Delta \).

By using this function, we define the inverse of \( \Lambda \). For any function \( f \) by a deep neural network on \( \Delta \), we define \( \Phi(f) = f \circ \text{Sort} \). We confirm \( \Lambda \circ \Phi = \text{id}_{\mathcal{F}_\Delta} \) and \( \Phi \circ \Lambda = \text{id}_{\mathcal{F}^S_n} \). Since we have
\[
\Lambda \circ \Phi(f) = \Lambda \circ f \circ \text{Sort} = (f \circ \text{Sort})_{\mid \Delta} = f,
\]
\( \Lambda \circ \Phi \) is equal to \( \text{id}_{\mathcal{F}_\Delta} \). Similarly,
\[
\Phi \circ \Lambda(f) = \Phi \circ f_{\mid \Delta} = f_{\mid \Delta} \circ \text{Sort} = f,
\]
where the last equality follows from the \( S_n \)-invariance of \( f \). Hence, we have the desired result.

Now, we are ready to prove Theorem 4.

Proof of Theorem 4 Let \( f^* \) be an \( S_n \)-invariant function on \( I \). Then by Proposition 9, we have a function \( f \) on \( \Delta_{S_n} \) such that \( f^* = f \circ \text{Sort} \) holds. By Theorem 5 in Schmidt-Hieber [2017], for enough big \( N \), there exists a constant \( c > 0 \) and a neural network \( f' \) with at most \( O(\log(N)) \) layers and at most \( O(N \log(N)) \) nonzero weights such that \( \| f - f' \|_{L^\infty(I)} \leq cN^{-\alpha/p} \). Then, we have
\[
\| f^* - f' \circ \text{Sort} \|_{L^\infty(I)} = \| f \circ \text{Sort} - f' \circ \text{Sort} \|_{L^\infty(I)} \leq \| f - f' \|_{L^\infty(\Delta)} \leq cN^{-\alpha/p},
\]
where \( f \circ \text{Sort} \) is a neural network with at most \( O(\log(N)) \) + \( K_1 \) layers and at most \( O(N \log(N)) \) + \( K_2 \) nonzero weights, where \( K_1 \) and \( K_2 \) are the number of layers and the number of nonzero weights of the neural network expressing \( \text{Sort} \) respectively. By replacing \( N^{-1} \) with \( \varepsilon \), we have the desired inequality.
B GENERALIZATION BOUND FOR EQUIVALENT DNN WITHOUT TRANSITIVE ASSUMPTION

In this section, we provide a general version of the result in Section 5. Namely, we relax the transitive assumption in the section. To the goal, we newly define a general version of a stabilizer subgroup.

Let \([n] = \{1, 2, \ldots, n\}\) be an index set and \(G\) be a finite group acting on \([n]\). For \(i \in [n]\), we define the stabilizer subgroup \(\text{Stab}_G(i)\) associated with \(G\) as

\[
\text{Stab}_G(i) = \{ \sigma \in G \mid \sigma \cdot i = i \}.
\]

We also consider the following decomposition of \([n]\) as

\[
[n] = \bigcup_{j \in J} \mathcal{O}_j,
\]

where \(J \subset I\) and \(\mathcal{O}_j\) is a \(G\)-orbit of \(j\), namely the set of the form \(G \cdot j\). Any \(G\)-orbit \(G \cdot j\) is isomorphic to the set \(G/\text{Stab}(j)\).

We denote \(|J|\) by \(J\) and \(|\mathcal{O}_j|\) by \(l_j\). For each \(j \in J\), let \(I = \bigcup_{j \in J} \bigcup_{k=1}^l \text{Stab}_G(j)\tau_{j,k}\) be the coset decomposition by \(\text{Stab}_G(j)\). Then, we may assume that \(\tau_{j,k} \in G\) satisfies \(\tau_{j,k}^{-1}(j) = j + k\).

Then, we provide another representation for equivariant functions from the following study.:  

**Proposition 10** (Representation for Equivariant Functions Sannai et al., 2019). A map \(F : \mathbb{R}^n \to \mathbb{R}^n\) is \(G\)-equivariant if and only if \(F\) can be represented by \(F = (f_1 \circ \tau_{1,1}, f_1 \circ \tau_{1,2}, \ldots, f_1 \circ \tau_{1,1}, f_2 \circ \tau_{2,1}, \ldots, f_j \circ \tau_{j,1})\) for some \(\text{Stab}_G(j)\)-invariant functions \(f_j : \mathbb{R}^n \to \mathbb{R}\). Here, \(\tau_{j,k} \in G\) is regarded as a linear map \(\mathbb{R}^n \to \mathbb{R}^n\).

**Proposition 11.** For any \(\varepsilon > 0\), we have

\[
\mathcal{N}_{\varepsilon, m}(\mathcal{F}^G(I)) \leq \prod_{I \in J} \mathcal{N}_{\varepsilon, m}(\mathcal{F}^G(\text{Stab}_G(j)(I_{j}))),
\]

where \(I_{j} = [0,1]^l\). Further, if \(G = S_n\),

\[
\mathcal{N}_{\varepsilon, m}(\mathcal{F}^G(I)) \leq \mathcal{N}_{\varepsilon, m}(\mathcal{F}^{S_{n-1}}(I)).
\]

**Proof of Proposition 11** We put \(N_j = \mathcal{N}_{\varepsilon, m}(\mathcal{F}^{\text{Stab}_G(j)}(I))\). For each \(j \in J\), by the definition of covering numbers, there exist \(f_j^{(1)}, \ldots, f_j^{(N_j)} \in \mathcal{F}^{\text{Stab}_G(j)}(I_{j})\) such that for any \(f' \in \mathcal{F}^{\text{Stab}_G(j)}(I_{j})\), there exists \(f_j^{(p)}\) satisfying \(\|f' - f_j^{(p)}\|_\infty < \varepsilon\).

With a tuple \((p_1, \ldots, p_j)\), we consider a map \(F_{p_1, \ldots, p_j} : I \to \mathbb{R}^n\) from \(\mathcal{F}^G(I)\) and claim that balls \(B_\varepsilon(F_{p_1, \ldots, p_j})\) give a covering set of \(\mathcal{F}(I)\). Put \(F_{p_1, \ldots, p_j} = (f_1^{(p_1)} \circ \tau_{1,1}, f_1^{(p_1)} \circ \tau_{1,2}, \ldots, f_1^{(p_1)} \circ \tau_{1,1}, f_2^{(p_2)} \circ \tau_{2,1}, \ldots, f_j^{(p_j)} \circ \tau_{j,1})\). Then \(F_{p_1, \ldots, p_j}\) is a \(G\)-equivariant map. Also, since \(\tau_{j,k}\) is a linear map by Proposition 10, we can represent \(\tau_{j,k}\) by DNNs. Hence, \(F_{p_1, \ldots, p_j} \in \mathcal{F}^G(I)\) holds.

Fix \(F' \in \mathcal{F}^G(I)\) arbitrary. We have the representation \(F' = (f_1' \circ \tau_{1,1}, f_1' \circ \tau_{1,2}, \ldots, f_1' \circ \tau_{1,1}, f_2' \circ \tau_{2,1}, \ldots, f_j' \circ \tau_{j,1})\) by Proposition 10. Then, we can find a corresponding \(F_{p_1, \ldots, p_j}\) such as

\[
\|F_{p_1, \ldots, p_j} - F'\|_{L^\infty(I)} = \max\{\|f_j^{(p_j)} \circ \tau_{j,k} - f_j' \circ \tau_{j,k}\|_\infty \mid 1 \leq k \leq l_j, 1 \leq p_j \leq N_j\} \\
\leq \varepsilon.
\]

Hence, we have the first statement.

In the case of \(S_n\), we have \(J = 1\) and \(\text{Stab}(1) \cong S_{n-1}\). This gives the second statement.

Then, we obtain the following general bound:

**Theorem 5** (Generalization of Equivariant DNN). Suppose \(\mathcal{F}^G \in \mathcal{F}^G(I)\) is uniformly bounded by 1. Then, for any \(\varepsilon > 0\), the following inequality holds with probability at least \(1 - 2\varepsilon\):

\[
R(\mathcal{F}^G) \leq R_m(\mathcal{F}^G) + \sqrt{\frac{c}{\sum_{j \notin J} |\text{Stab}_G(j)| m^2/n}} + \sqrt{\frac{2 \log(2/\varepsilon)}{m}}.
\]
where $\tilde{c} > 0$ is a constant which are independent of $n$ and $m$.

We omit rigorous proof of Theorem 5, because it is almost same to that of Theorem 3.